REPORT

통계기반 데이터 분석

회귀분석을 활용한
인간의 기대 수명 예측에 대하여
- stat.x77 데이터셋을 바탕으로

2023.03.20

B1 팀

김예지, 서영석, 이현빈, 전국림

<목차>

1. 서론

1) 데이터 분석 배경

수명 분석으로 인간의 기대수명을 예측, 추론 하는 것에 대한 니즈(Needs)는 시대와 세대, 국경을 초월하여 존재해 왔다. 결과값을 요구하는 목적에는 의료, 경제 등의 활용 분야에 따라 다양하게 존재할 것이다. 이러한 범사회적인 니즈에 맞추어 이 보고서에서는 인간의 기대수명에 대한 분석을 실시할 것이다. 특히 '기대수명에 가장 큰 영향을 미치는 요인'에 대해 파악하고자 한다. 이를 통해 순영향을 파악하여, 결과적으로 인류의 기대수명 증가를 위한 데이터적 근거 기반을 형성하고자 한다.

2) 데이터 분석 설명

이 보고서에서는 R을 활용하여 인간의 기대 수명에 대한 데이터를 분석할 예정이다. 데이터셋은 일관성을 위해 R의 내장 데이터셋인 state data sets 내 stat.x77 데이터셋으로 제한한다.

해당 데이터는 미국 50개 주라는 지역을 표본으로 한다. 데이터셋 상의 변수명의 의미를 서론에서 먼저 밝혀, 보고서 내용 및 결과에 대한 이해를 돕고자 한다

변수	변수 정의
df_state	원본 데이터 프레임
Populationdesc	인구수 기준 내림차순 정렬 데이터 프레임
Life.Exp	기대수명
Population	미국 주 인구수(1975 년 1 월)
Income	인당 소득(1974 년)
Illiteracy	문맹률(1970 년)
Murder	인구 100,000 명당 살인율(1976 년)
HS.Grad	고등학교 졸업자 비율(1970 년)
Frost	최저 기온이 영하 미만 평균 일수(1931~1960 년)
Area	평방 마일의 토지 면적

3) 데이터 정제

본론에 들어가기에 앞서, 데이터셋을 가져오는 과정에 대해 먼저 밝히고자 한다. 먼저 "data(state)"코드를 통해 state 데이터셋을 로드(load)한다. 이후 중심적으로 활용할 데이터인, state 데이터셋에 있는 state.x77 dataset 을 데이터프레임으로 변환한다. 마지막으로 칼럼명에 대한 명확성을 높이기 위해 띄워쓰기가 되어있는 부분을 온점('.')으로 변환한다.

해당 부분에 대한 코드는 아래와 동일하며, 아래에서는 데이터셋의 정보를 'df_state'라는 이름의 변수에 담았다. 또한, 데이터에 대한 기본적인 이해를 돕기 위해 기술 통계량 및 데이터 구성을 summary 함수와 dim 함수를 통해 확인하였다.

```
# (1)-1 state dataset load
data(state) # -> state : 미국 1970 년대 50 개 주에 대한 데이터셋
?state  #데이터 셋에 대한 정보를 확인
# (1)-2 state.x77 dataset -> data frame 으로 변환
df_state <- as.data.frame(state.x77)
# (1)-3 Life Exp 변수 -> Life.Exp 변경
colnames(df_state)[colnames(df_state)=="Life Exp"] <- "Life.Exp"
# (1)-4 HS Grad 변수 -> HS.Grad 변경
colnames(df_state)[colnames(df_state)=="HS Grad"] <- "HS.Grad"
# 기술통계량 확인용
summary(df_state)
# 데이터 구성 확인용
dim(df_state) # 50 개 주 / 8 개 특성
```

df_state 에 담긴 데이터에는 총 50 개 주의 정보가 담겨져 있으며, Income, Murder 등 8 개의 특성을 칼럼으로 담고 있다.

2. 본론

1) Life Expectancy 변수 중심의 데이터 분석

이 보고서에서는 인간의 기대수명에 대하여 확인하고자 한다. 따라서 Life Expectancy 변수를 종속변수로 설정하고 나머지 변수를 독립변수로 설정했다. 또한, 회귀분석을 통해 결과값을 예측하고자 한다. ('Population' 칼럼의 데이터는 "실제 인구수/1000"으로 기재 되어 있다.)

외부 요인에 따른 기대수명의 변화를 확인하기 위해서는 모든 요인을 포함한 분석값이 존재해야 한다. 따라서 Life Expectancy 변수를 중심으로 데이터셋에 존재하는 전체 요인(칼럼)에 대한 회귀분석을, lm 함수를 활용해 실시했다.

```
result.lm <- lm(Life.Exp ~., data = df_state)
result.lm
summary(result.lm)
```

위의 코드에 대한 결과 값은 아래와 동일하게 출력된다.

> summary(result.lm)

Call :

lm(formula = Life.Exp ~ ., data = df_state)

Residuals:

Min	1Q	Median	3Q	Max
-1.48895	-0.51232	-0.02747	0.57002	1.49447

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.94322411113	1.74797537818	40.586	< 0.0000000000000002
Population	0.00005180036	0.00002918703	1.775	0.0832
Income	-0.00002180424	0.00024442561	-0.089	0.9293
Illiteracy	0.03382032136	0.36627989117	0.092	0.9269
Murder	-0.30112317045	0.04662072985	-6.459	0.0000000868
HS.Grad	0.04892947888	0.02332327770	2.098	0.0420
Frost	-0.00573500110	0.00314322966	-1.825	0.0752
Area	-0.00000007383	0.00000166816	-0.044	0.9649

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7448 on 42 degrees of freedom

Multiple R-squared: 0.7362, Adjusted R-squared: 0.6922

F-statistic: 16.74 on 7 and 42 DF, p-value: 0.0000000002534

해당 결과에 대한 이해를 돕기 위해 해석을 추가한다.

Intercept (절편)은 70.94 이다. 즉 모든 변수가 0 일 때의 기대 수명은 70.94 살이다.

Population(인구수)은 p-value 값이 유의계수 0.05 보다 커 통계가 유의하지 않다.

Income(인당 소득), Illiteracy(문맹률), Frost(최저 기온이 영하 미만인 일수), Area(토지 면적) 또한 p-value 값이 0.9269 로 유의계수 0.05 보다 커 통계가 유의하지 않다.

반면 Murder(인구 100,000 명당 살인율)의 경우, p-value 값이 0.05 보다 작아 통계가 유의하다. 여기서는 살인율의 계수가 -0.3011 으로 살인율이 1 증가할때 마다 기대수명이 0.3011 씩 줄어든다는 것을 예측할 수 있다.

또한, HS.Grad(고졸 비율)은 p-value 값이 0.04893 로 0.05 보다 작아 통계가 유의하다. 즉 학력이 고등학교 졸업 수준일 때의 비율이 1 증가할 때 마다, 기대수명이 0.04893 늘어남을 예측할수있다.

이어서 다중 공선성에 대한 분석도 실시했다. 다중 공선성이란 독립 변수들 간의 상관정도이다. 상관정도가 높을수록, 회귀 분석에서는 분석 결과인 회귀 계수가 불안정해 진다. 즉 회귀계수가 해당 변수의 종속변수에 미치는 영향력을 올바로 설명하지 못하게 된다. 이러한 잘못된 결과를 도출하지 않기 위해 다중 공선성에 대한 분석이 필요하다. 이를 위한 코드는 아래와 동일하다.

```
# 패키지 설치
# install.packages("car")
library(car)

# 분산팽창요인(VIF)
vif(result.lm)
```

위의 코드를 실행한 결과 아래의 도표와 동일한 결과값이 출력된다. 일반적으로 다중공선성은 10 이상일 경우 의심을 하며, 엄격한 기준을 적용했을 경우에는 4 이상으로 설정한다.

Population	Income	Illiteracy	Murder	HS.Grad	Frost	Area
1.499915	1.992680	4.403151	2.616472	3.134887	2.358206	1.789764

실행 결과 모든 변수의 다중공선성 값이 모두 1~5 사이이다. 따라서 해당 변수들은 다중 공선성에 대한 문제가 존재하지 않는 것으로 판단할 수 있다.

2) 변수 한정을 통한 기대 수명 분석

이어서 df_state 변수 한정에 따른 기대수명의 변화에 대한 결과를 도출하고자 한다. 따라서 Income, Illiteracy, Area 변수를 제외한 한정된 회귀 분석을 실시했다. 동시에 다중 공선성에 대한 분석도 실시했으며, 해당 코드는 아래와 동일하다.

```
func3 <- Life.Exp ~Population+Murder+HS.Grad+Frost
result2.lm <- lm(func3, data = df_state)

result2.lm
summary(result2.lm)
vif(result2.lm)
```

위의 코드에 대한 결과 값은 아래 사진과 동일하게 출력된다.

```
> summary(result2.lm)
Call :
lm(formula = func3, data = df_state)
Residuals:
```

Min	1Q	Median	3Q	Max
-1.47095	-0.53464	-0.03701	0.57621	1.50683

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.02712853	0.95285296	74.542	< 0.0000000000000002
Population	0.00005014	0.00002512	1.996	0.05201
Murder	-0.30014880	0.03660946	-8.199	0.000000000177
HS.Grad	0.04658225	0.01482706	3.142	0.00297
Frost	-0.00594329	0.00242087	-2.455	0.01802

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 0.000000000001696
```

이러한 코드 결과를 통해 기대수명을 분석할 수 있다. 결과에서 나온 수치 중 'Pr(>|t|)'은 p-value 값이다.

먼저 Intercept (절편)은 71.03 으로, 모든 변수가 0 일 때, 기대 수명이 71.03 살임을 의미한다. 변수를 한정하기 전보다 기대수명이 증가했다.

Population(인구수)는 p-value 값이 0.05201 므로 유의계수 0.05 보다 커, 통계가 유의하지 않다.

반면, Murder(인구 100,000 명당 살인율)은 p-value 값이 0.05 보다 작아 통계가 유의하다. 수치상으로 살인율의 계수가 -0.3001 이다. 즉, 살인율이 1 증가할때 마다 기대 수명이 0.3001 줄어든다는 것을 예측할 수 있다.

HS.Grad(고졸 비율)의 p-value 값은 0.00297 로, 0.05 보다 작아 통계가 유의하다. 고졸 비율 계수가 0.046 이므로, 고졸 비율이 1 증가할때 마다 기대수명이 0.046 늘어남을 예측할 수 있다.

또한, Frost(최저 기온이 영하 미만인 평균 일수)의 p-value 값은 0.01802 이며, p-value 값이 0.05 보다 작아 유의하다. 이로써 영하인 일수가 많을 수록 기대수명이 미세하게 줄어듬을 예측할 수 있다.

Population	Murder	HS.Grad	Frost
1.189835	1.727844	1.356791	1.498077

이러한 결과에서 다중 공선성은 위의 표와 동일하다. 여기서 다중 공선성은 모두 1~2 사이에 존재하므로, 모든 값이 문제가 존재하지 않다고 판단할 수 있다.

3) 예측 변수를 활용한 기대수명 분석

3-1) HS.Grad 와 Murder 변수를 예측변수로 설정한 회귀분석

여기서는 HS.Grad, Murder 에 해당하는 변수만을 포함한 한정된 회귀 분석을 실시했다. 동시에 다중 공선성에 대한 분석도 vif() 함수를 활용해 실시했으며, 해당 코드는 아래와 동일하다.

```
func4 <- Life.Exp ~HS.Grad+Murder
result3.lm <- lm(func4, data = df_state)
result3.lm
vif(result3.lm)
```

위의 코드에 대한 결과 값은 아래와 동일하게 출력된다.

```
> result3.lm
Call :
lm(formula = func4, data = df_state)
```

(Intercept)	HS.Grad	Murder
70.29708	0. 0.04389	-0.23709

여기서는 Intercept (절편)이 70.29708 이다. 모든 변수가 0 일 때 기대 수명이 70.29708 살임을 의미한다.

또한 HS.Grad (고졸)의 p-value 값은 0.04389 로, 고등학교 졸업율이 증가할수록 기대 수명은 0.04389 증가한다.

Murder (살인 발생률)의 p-value 값은 -0.23709 이며, 살인 발생률이 증가할수록 기대 수명은 0.23709 감소함을 의미한다.

3-2) 예측 변수로 설정한 회귀 분석

3-1)에서 나온 분석 값을 기준으로 아래와 동일한 가정을 추가하여, 구체적인 예측을 실시하고자 한다.

가정) 전 인구의 55%가 고졸이고 살인비율이 10 만명당 8 명일 때 Life Expectancy

따라서 3-1)에서의 회귀분석 변수인 result3.lm 에서 구한 절편과 기울기를 사용하여 회귀 방정식 만들고자 한다.

회귀 방정식은 "Y = 70.29708 -0.23709*(Murder) + 0.04389*(HS.Grad)"이라고 할 수 있다.

또한, 가정에서 전 인구의 55%가 고졸이라고 했으므로, 고졸비율(HS.Grad)은 55.0 이 된다. 살인비율이 10 만명당 8 명이기에, 살인비율(Murder)에는 8.0 를 대입한다.

결과적으로 "Y = 70.29708 + (-0.23709*8.0) + (0.04389*55.0)"의 식이 완성된다.

```
Y = 70.29708 + (-0.23709*8.0) + (0.04389*55.0)
Y
```

```
> Y
[1] 70.81431
```

따라서 해당 가정에서의 기대수명 예측 값은 70.81431 이라고 할 수 있다.

3. 분석 결과 및 결론

1) 데이터 분석 결과 및 결론 도출

본론 3 에서 확인할 수 있는 회귀 값을 정리하고자 한다. 동시에 본 보고서의 목적인 '기대수명에 높은 영향을 미치는 요인' 파악의 용이성을 높이고자 한다. 먼저 기대 수명에 높은 영향을 미치는 요인에 대해 분석했다. 이를 위한 코드는 아래와 동일하다.

```
# 1) 회귀식 생성하기
result7.lm <- lm(Life.Exp ~., data = df_state)
fun <- lm(result7.lm, data = df_state,family = 'binomial', na.action=na.omit)
# 2) 후진제거법을 활용한 변수 선택
select <- step(fun, direction = "backward")
```

이를 통한 최종 결과는 아래 사진과 동일하다. 즉 인구, 살인비율, 고졸율, 최저 기온이 영하 미만인 평균 일수가 높은 영향을 미치는 것으로 확인할 수 있다.

Step : AIC = -28.16

Life.EXP ~Population + Murder + HS.Grad + Frost

	Df	Sum of Sq	RSS	AIC
<none>			23.308	-28.161
Population	1	2.064	25.372	-25.920
Frost	1	3.122	26.430	-23.877
HS.Grad	1	5.112	28.420	-20.246
Murder	1	34.816	58.124	15.528

마지막으로 2 개의 독립변수(HS.Grad, Murder)와 1 개의 종속변수(Life.Exp)의 데이터와 fit 된 회귀평면을 3D 그래프로 시각화하여 표현했다. 코드는 아래와 동일하다.

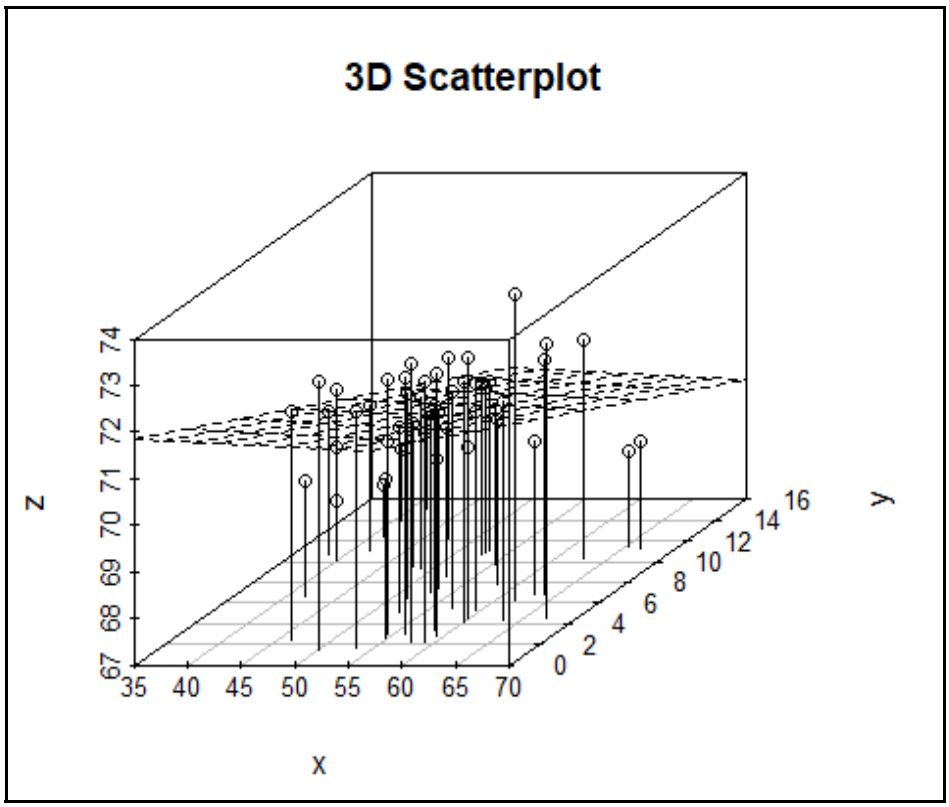
```
library(scatterplot3d)
x <- df_state$HS.Grad
y <- df_state$Murder
z <- df_state$Life.Exp

# 회귀 평면 구하기
fit <- lm(z ~ x + y)

# 3D scatter plot 생성
s3d <- scatterplot3d(x, y, z, type="h", main="3D Scatterplot")

# 회귀 평면 추가
s3d$plane3d(fit)
```

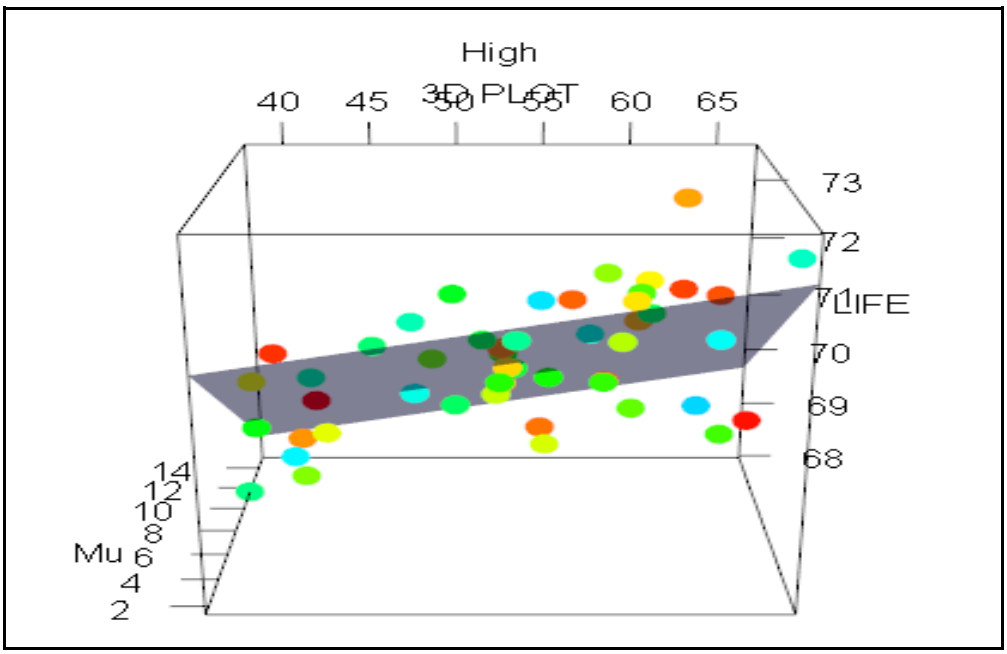
코드를 통해 얻을 수 있는 시각화 자료는 아래와 동일하다. 해당 3D 그래프에서, x 축은 고졸 비율을, y 축은 살인 비율을, z 축은 기대 수명을 의미한다.



위의 그래프보다 더욱 입체적인 그래프는 아래 코드를 통해 확인할 수 있다.

```
library(plot3Drgl)
library(rgl)
library(plot3D)
# 데이터 생성
x1 <- df_state$HS.Grad
x2 <- df_state$Murder
y1 <- df_state$Life.Exp
# 회귀분석 모델 적합
fit <- lm(y1 ~ x1 + x2)
# 회귀분석 평면 시각화
#산포도시각화
rainbowcolor <- rainbow(93) #레인보우 색상 랜덤
plot3d(x = x1, y = x2, z = y1, type = "p",size = 13,   col = rainbowcolor ,xlab = "High", ylab = "Mu", zlab = "LIFE", main = "3D PLOT")
#회귀평면 추가
b0 <- coef(fit)[1]#회귀계수 담기
b1 <- coef(fit)[2]# 고졸
b2 <- coef(fit)[3]#  살인
x1.grid <- seq(min(x1), max(x1), length.out = 10) #x 축면의 간격 지정
x2.grid <- seq(min(x2), max(x2), length.out = 10) #y 축면의 간격 지정
y1.grid <- outer(x1.grid, x2.grid, function(x1, x2) b0 + b1*x1 + b2*x2) #z 축면 간격 지정
rgl.planes(a = -b1, b = -b2, c = 1, d = -b0, alpha = 0.5, col = "blue") #alpha 투명도
```

코드를 통해 얻을 수 있는 시각화 자료는 아래와 동일하다. 해당 3D 그래프에서, High 는 고졸 비율을, Mu 는 살인 비율을, Life 는 기대 수명을 의미한다.



참고 자료

1) 데이터셋: state data sets 내 stat.x77 데이터셋