Multi_linear_regression

juhwan

5/26/2021

서론

혁신과 관련하여 다중 회귀 분석을 할 것이다.
- 종속변수 = Global Innovation Index
- 독립변수 = 소속 대륙, 문화적 차원 점수, 이외의 1가지

본론

라이브러리 불러오기

```
suppressMessages(library(magrittr))
suppressMessages(library(dplyr)) %>% suppressWarnings()
suppressMessages(library(jtools)) %>% suppressWarnings()
suppressMessages(library(car)) %>% suppressWarnings()
suppressMessages(library(mice)) %>% suppressWarnings()
suppressMessages(library(sjPlot)) %>% suppressWarnings()
suppressMessages(library(QuantPsyc)) %>% suppressWarnings()
suppressMessages(library(lmtest)) %>% suppressWarnings()
```

파일 불러오기

```
df_culture = read.csv('6-dimensions-for-website-2015-08-16.csv', header = T, sep =
';')
df_inno = read.csv('Analysis_2021may.24161011.csv', header = T)
df_country = read.csv('countryContinent.csv', header = T)
# pdi = 권력격차, idv = 개인주의, mas = 남여 불균형, uai = 불확실성 회피지수, ltowvs = 장기지향성
, ivr = 쾌락 추구
```

전처리

```
head(df_culture) # 대략적 확인
```

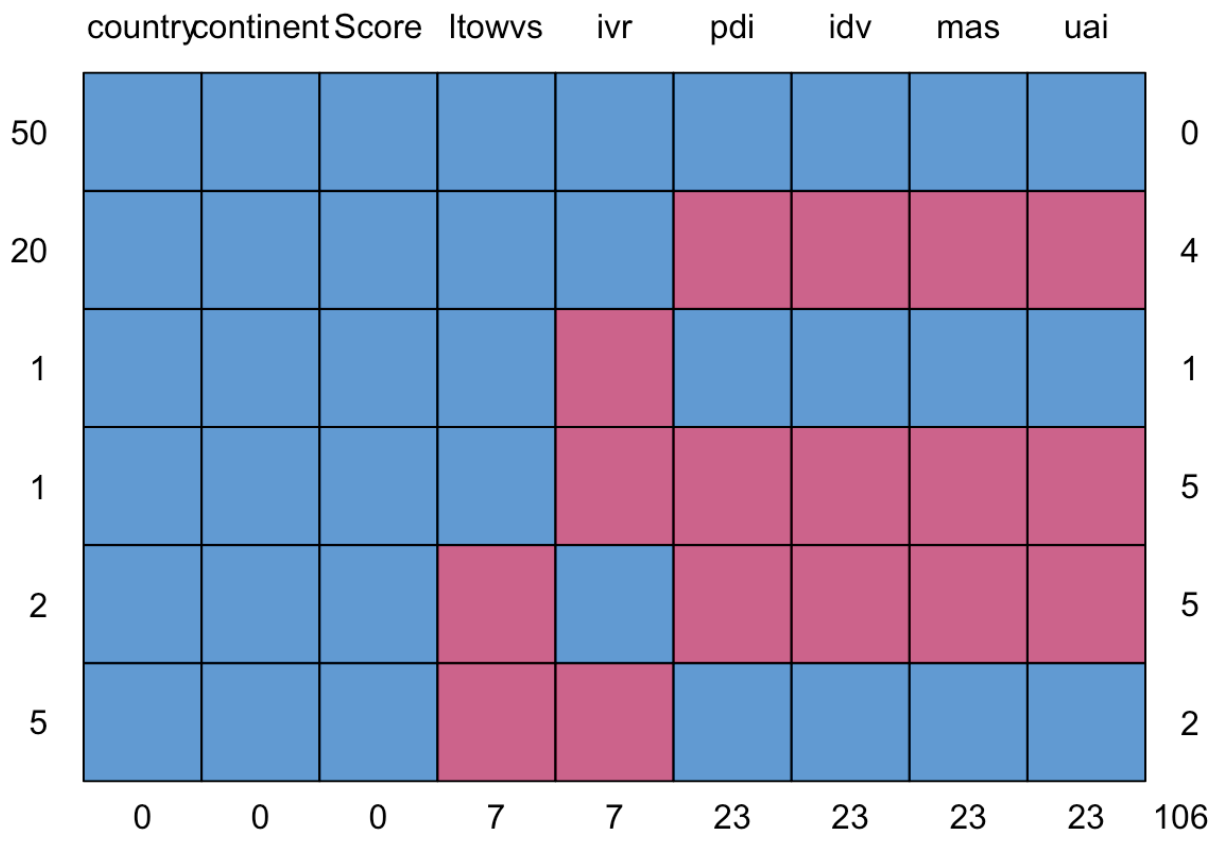
```
##   ctr        country    pdi    idv    mas    uai ltowvs ivr
## 1 AFE    Africa East     64     27     41     52     32  40
## 2 AFW    Africa West     77     20     46     54      9  78
## 3 ALB        Albania #NULL! #NULL! #NULL! #NULL!     61  15
## 4 ALG        Algeria #NULL! #NULL! #NULL! #NULL!     26  32
## 5 AND        Andorra #NULL! #NULL! #NULL! #NULL! #NULL!  65
## 6 ARA Arab countries     80     38     53     68     23  34
```

```
df_culture[] = lapply(df_culture, function(x) gsub("#NULL!", NA, x)) # '#NULL!'값 제거
df_culture = df_culture[, -1] # 필요 행만 서브셋팅
df_inno = df_inno[,c(2, 5)] # 필요 행만 서브세팅
df_country = df_country[, c(1, 6)] # 필요 행만 서브세팅
colnames(df_inno) = c('country', 'Score') # 열 들의 이름 바꾸기

df_inno$Score = as.numeric(df_inno$Score) # 열 숫자로 바꾸기
rownames(df_culture) = df_culture$country ; df_culture = df_culture[,-1] # 나라 이름을 행으로 변환
df_culture[] = lapply(df_culture, function(x) as.numeric(x)) # 데이터 프레임 숫자로 바꾸기
df_culture$country = rownames(df_culture)

## 병합
df = inner_join(df_culture, df_country, by = 'country')
df = inner_join(df, df_inno, by = 'country')

## na 제거
md.pattern(df) # 결측치에 대한 통계를 본다.
```




```
##    country continent Score ltowvs ivr pdi idv mas uai
## 50       1         1     1      1   1   1   1   1   0
## 20       1         1     1      1   1   0   0   0   4
## 1        1         1     1      1   0   1   1   1   1
## 1        1         1     1      1   0   0   0   0   5
## 2        1         1     1      0   1   0   0   0   5
## 5        1         1     1      0   0   1   1   1   2
##          0         0     0      7   7  23  23  23  23 106
```

```
df = na.omit(df) # NA값이 잇는 모든 행을 삭제한다.

## 더미변수 생성
factor(df$continent) # 팩터형을 통해 5 대륙 중 빠진 것이 있나 확인
```

```
##  [1] Americas Oceania  Europe   Asia     Europe   Americas Europe   Americas
##  [9] Americas Asia     Americas Europe   Europe   Americas Europe   Europe
## [17] Europe   Europe   Europe   Europe   Asia     Asia     Europe   Europe
## [25] Asia     Europe   Europe   Europe   Asia     Europe   Americas Africa
## [33] Oceania  Europe   Asia     Americas Asia     Europe   Europe   Europe
## [41] Europe   Asia     Europe   Europe   Europe   Europe   Asia     Americas
## [49] Asia     Americas
## Levels: Africa Americas Asia Europe Oceania
```

```
# 모든 대륙이 하나 이상은 있다. 따라서 4개의 더미변수 생성 작업
df$Africa = ifelse(df$continent == 'Africa', 1, 0)   # 아프리카 생성
df$Americas = ifelse(df$continent == 'Americas', 1, 0)  # 아메리카 생성
df$europe = ifelse(df$continent == 'Europe', 1, 0)  # 유럽 생성
df$Oceania = ifelse(df$continent == 'Oceania', 1, 0)  # 오세아니아 생성
# 준거집단은 아시아, 즉 위의 네가지가 0이면 아시아
```

선형 회귀

```
multi_linear_regression = lm(Score ~ pdi + idv + mas + uai + ltowvs + ivr + europe
+ Americas + Oceania + Africa, data = df)
summary(multi_linear_regression)
```

```
##
## Call:
## lm(formula = Score ~ pdi + idv + mas + uai + ltowvs + ivr + europe +
##     Americas + Oceania + Africa, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2463  -2.3693  -0.1042   2.4031  12.9500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.27705    7.22514   2.945 0.005423 **
## pdi         -0.01303    0.06037  -0.216 0.830290
## idv          0.13832    0.06319   2.189 0.034645 *
## mas          0.00243    0.04880   0.050 0.960540
## uai         -0.12564    0.04850  -2.591 0.013406 *
## ltowvs       0.21312    0.05772   3.692 0.000679 ***
## ivr          0.24353    0.05537   4.398 8.18e-05 ***
## europe       5.34650    3.42897   1.559 0.127024
## Americas    -5.62660    3.85115  -1.461 0.152022
## Oceania     -2.26867    5.98128  -0.379 0.706526
## Africa       1.61510    6.79983   0.238 0.813496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.768 on 39 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.726
## F-statistic: 13.98 on 10 and 39 DF,  p-value: 4.855e-10
```

```
# 표준화 계수값 구하기
lm.beta(multi_linear_regression) # 가장 영향을 많이 미치는 변수는 ivr이다.
```

```
##          pdi          idv          mas          uai       ltowvs          ivr
## -0.024241112  0.283671675  0.004290035 -0.264730740  0.406362865  0.483828322
##       europe     Americas      Oceania       Africa
##  0.244842699 -0.206300754 -0.040750422  0.020726306
```

모형선택 : 전진 선택 방법

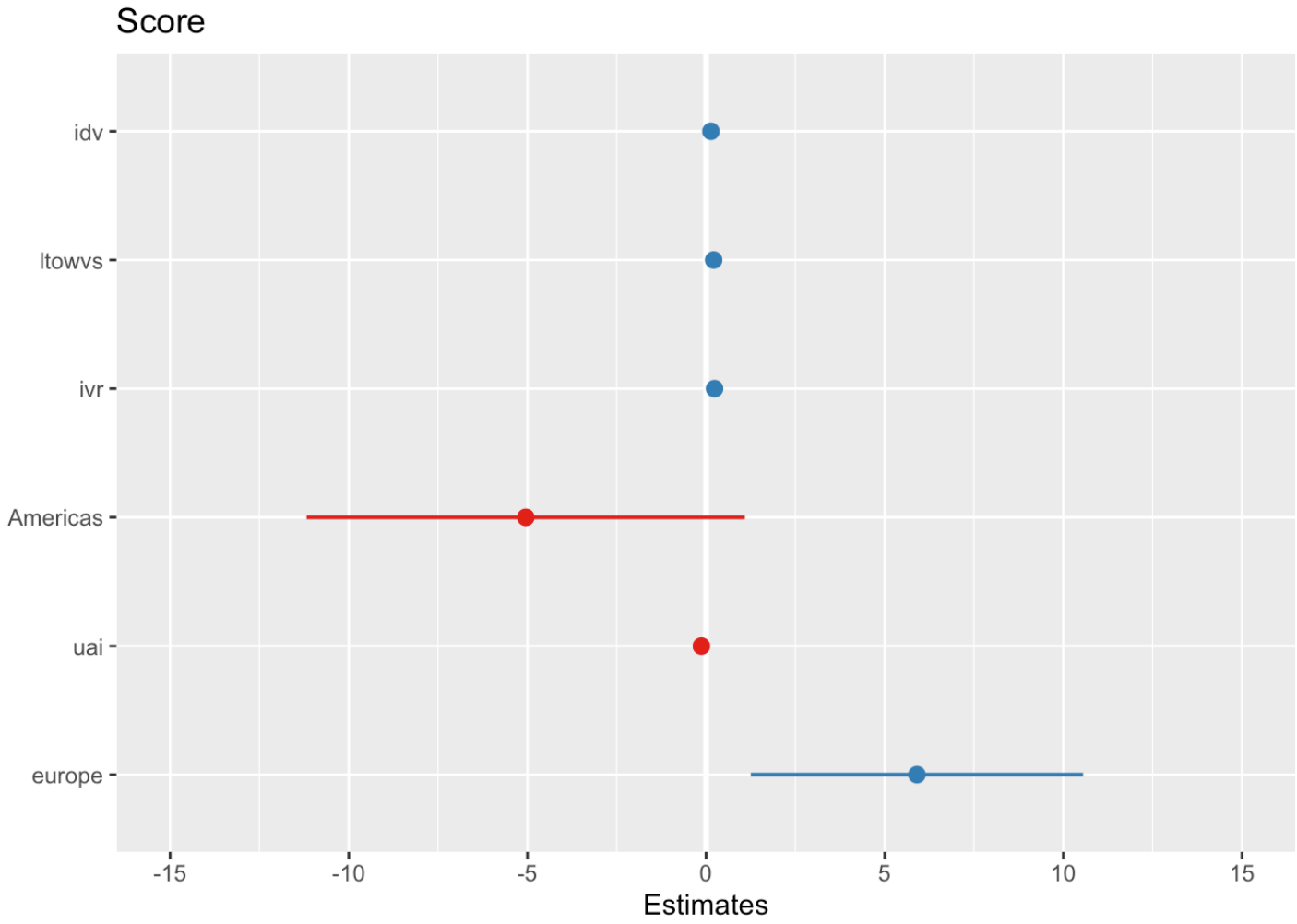
```
null = lm(Score ~ 1, data = df)
front_model = step(null, scope=list(lower=null, upper=multi_linear_regression), di
rection="forward") # 실행은 시켰지만, 너무 길어지기 때문에 result = 'hide' 옵션을 통해 콘솔 출력
을 막았다.
```

검증

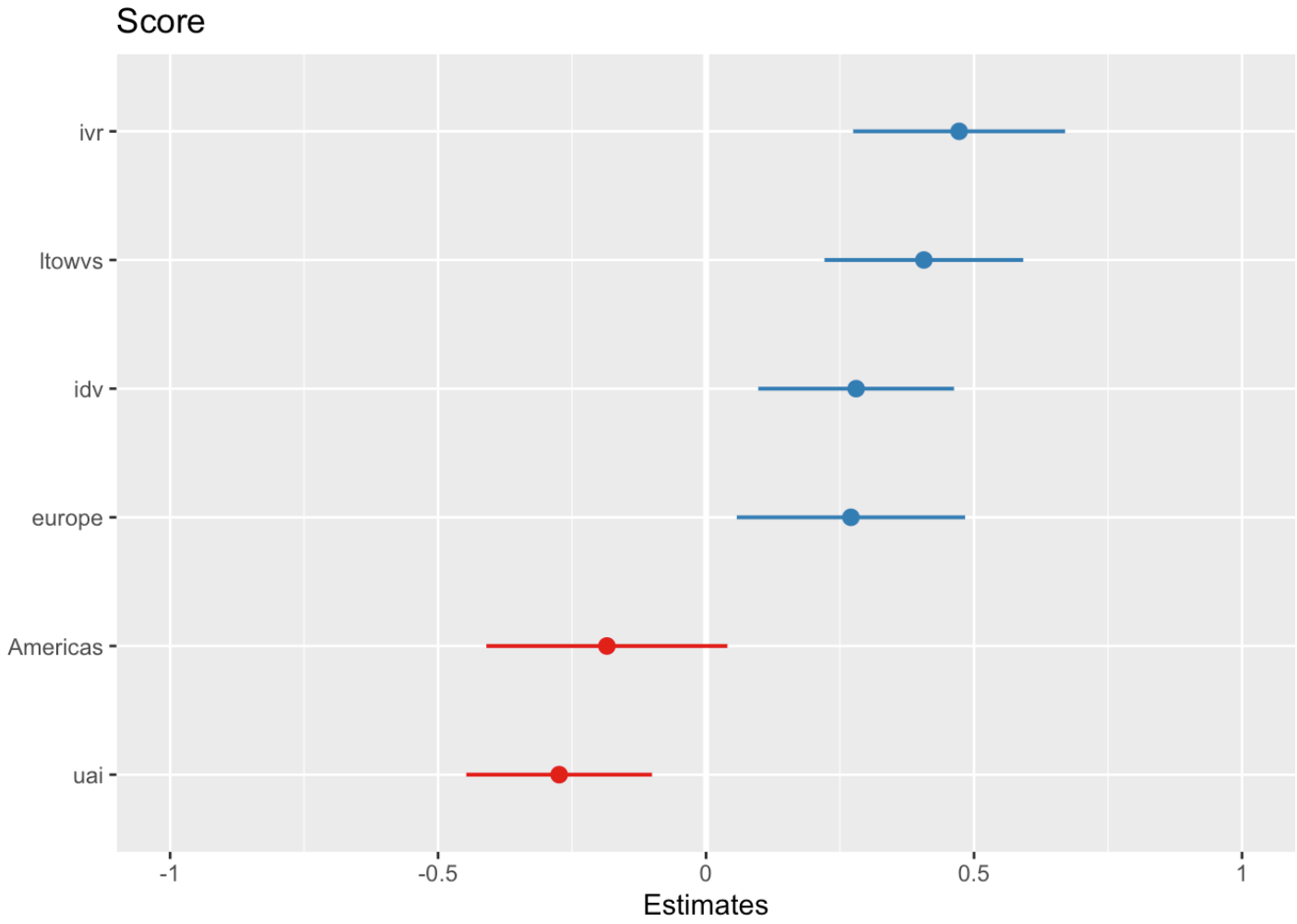
```
tab_model(front_model) # 깔끔하게 본다. 결정계수는 0.78
```

	Score		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	20.85	10.95 – 30.75	**<0.001**
idv	0.14	0.05 – 0.23	**0.003**
ltowvs	0.21	0.12 – 0.31	**<0.001**
ivr	0.24	0.14 – 0.34	**<0.001**
Americas	-5.05	-11.18 – 1.09	0.104
uai	-0.13	-0.21 – -0.05	**0.003**
europe	5.90	1.25 – 10.55	**0.014**
Observations	50		
R^2 / R^2 adjusted	0.780 / 0.750		

```
plot_model(front_model, type = "est", wrap.labels=5) # 회귀 계수를 살펴본다.
```




```
plot_model(front_model, type = "std", sort.est = T, wrap.labels=5) # 표준화 계수를 살펴본다.
```

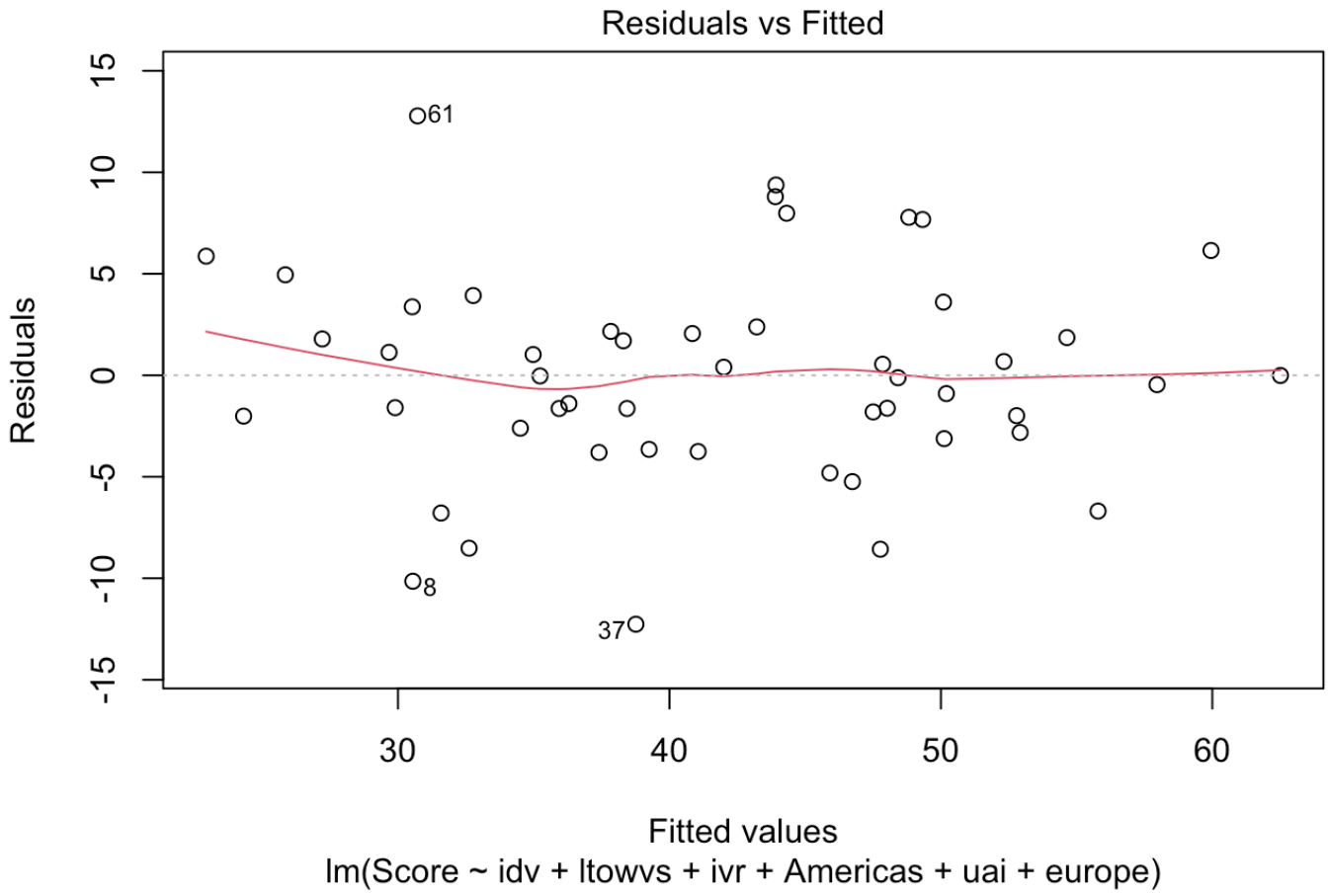



```
# 다중공산성 확인
vif(front_model)
```

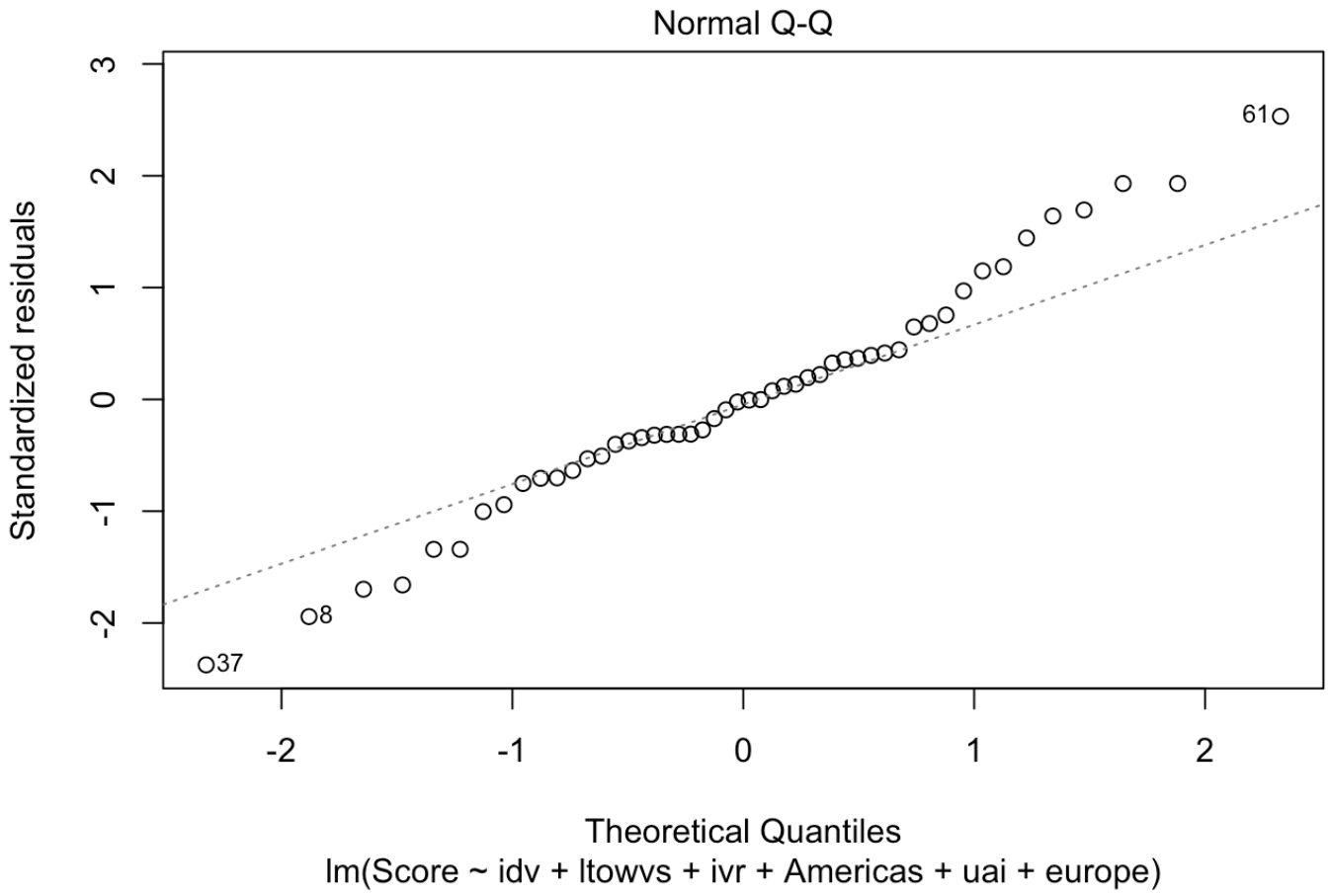
```
##      idv   ltowvs      ivr Americas      uai   europe
## 1.605360 1.656077 1.881493 2.433857 1.445581 2.183030
```

```
# 모두 괜춘 5 이하.

plot(front_model, 1) # 모형의 선형성이 인정됨.
```



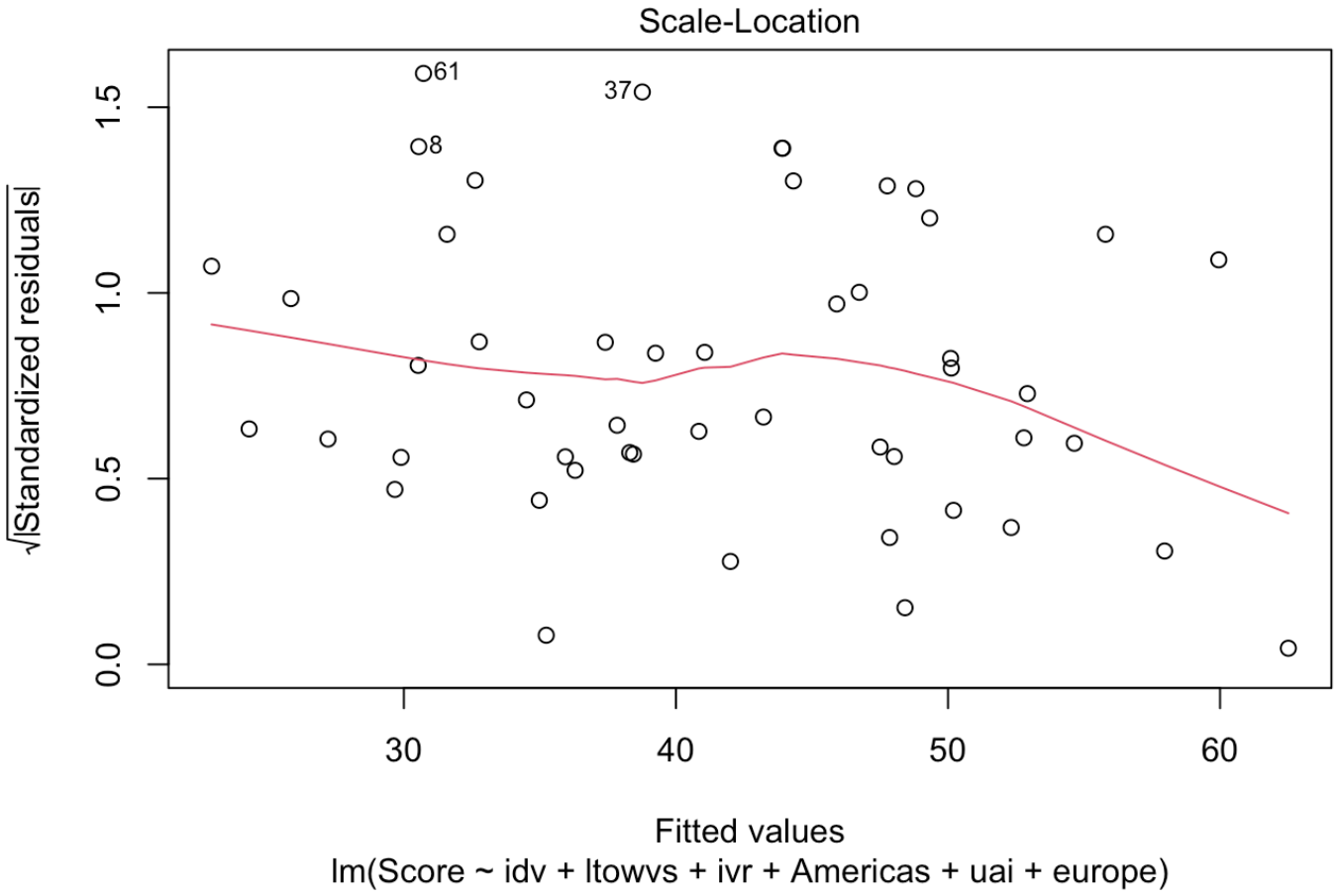

```
plot(front_model, 2) # 잔차의 정규성이 검증됨.
```



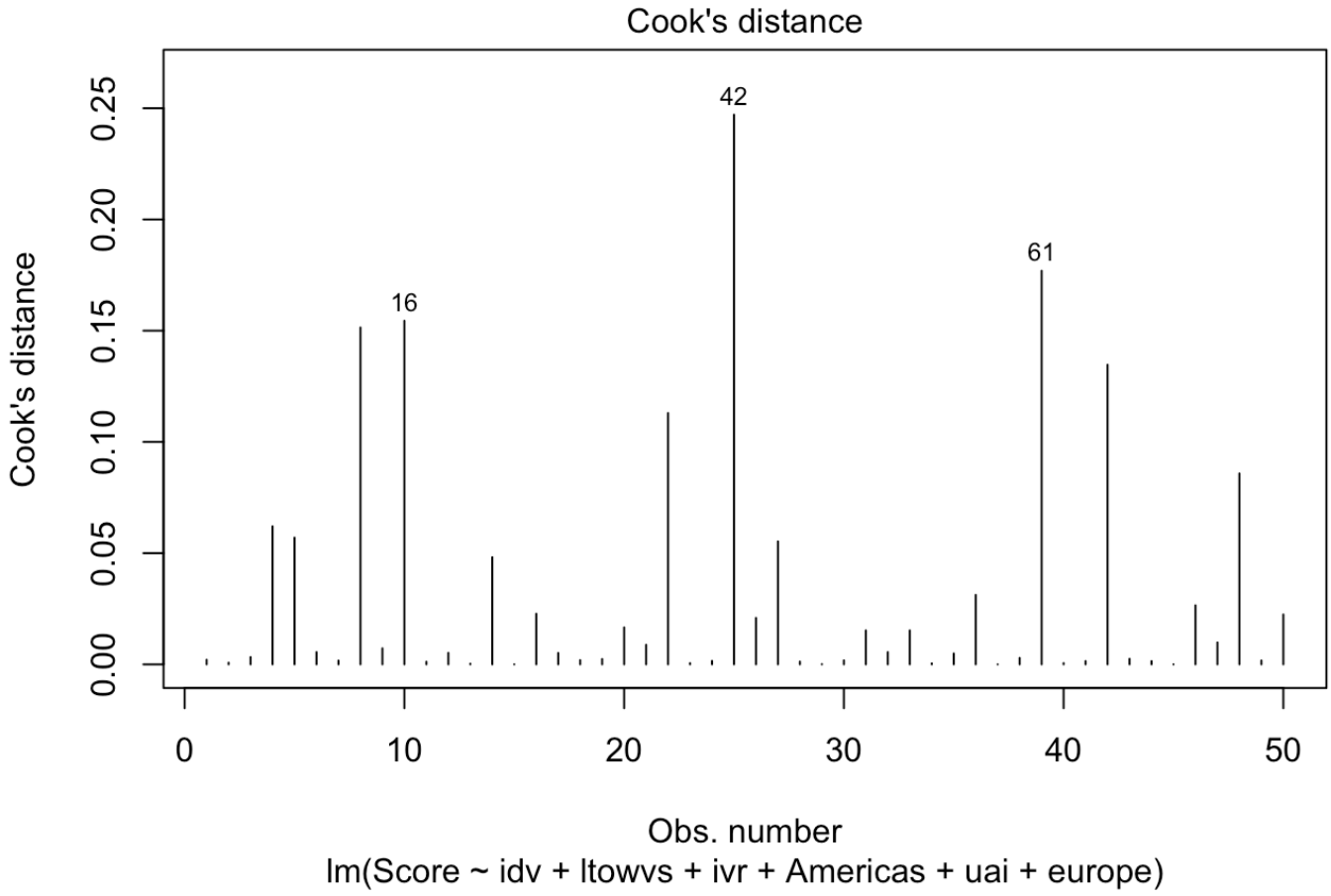

```
shapiro.test(front_model$residuals) # 샤피로 테스트로 p값이 0.87이니 정규성 가정 만족.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  front_model$residuals
## W = 0.98766, p-value = 0.8771
```

```
plot(front_model, 3) # 약간 아쉽지만 잔차의 등분산성이 만족.
```




```
plot(front_model, 4) # 16번, 42번, 61번의 자료가 예측치에 비해 특히 벗이난다.
```




```
dwtest(front_model) # 자기상관 계수가 1.5이니 자기 상관이 없는 것으로 검증완료.
```

```
##
##  Durbin-Watson test
##
## data:  front_model
## DW = 1.5017, p-value = 0.03152
## alternative hypothesis: true autocorrelation is greater than 0
```

검증결과 위의 분석이 모두 유의미한 것으로 검증이 완료되었다.

변수 추가

```
# 특정 변수를 추가해 회귀선의 R^2값을 키워볼 것이다.
data_3rd = read.csv("population_by_country_2020.csv") # 데이터 불러오기
data_3rd = data_3rd[, c(1, 9)] # 국가 이름과 국가별 나이의 중간 값을 가져온다.
colnames(data_3rd) = c("country", "mid_age")
df_plus = inner_join(df, data_3rd, by = 'country') # 병합

#회귀 다시 구하기
df_plus = df_plus[, -7]
df_plus$mid_age = as.numeric(df_plus$mid_age)
null_plus = lm(Score ~ 1, data = df_plus)
multi_linear_regression_plus = lm(Score ~ ., data = df_plus)
front_model_plus = step(null_plus, scope=list(lower=null_plus, upper=multi_linear_
regression_plus), direction="forward")
```

```
tab_model(front_model_plus) # 결정계수 값 증가.
```

	Score		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.68	-6.18 – 13.53	0.456
idv	0.13	0.05 – 0.21	**0.001**
mid_age	0.72	0.42 – 1.01	**<0.001**
uai	-0.15	-0.22 – -0.09	**<0.001**
ivr	0.22	0.13 – 0.30	**<0.001**
ltowvs	0.13	0.04 – 0.22	**0.008**
Americas	-6.08	-11.03 – -1.13	**0.017**
Oceania	-5.94	-13.79 – 1.90	0.134
Observations	50		
R^2 / R^2 adjusted	0.848 / 0.823		

결론

위의 다중 선형 회귀 결과
1. F값을 통해 위의 선형 회귀가 데이터를 설명할 수 있다는 대립가설 채택.
2. 유의미한 독립변수로 ltowvs와 ivr는 매우 유의미함.
3. 즉, 장기지향성과 쾌락을 추구할 수록 국가적 혁신 지수가 높아짐을 알 수 있다.
4. 위의 회귀선은 전체 데이터의 약 78%를 설명함.
5. 또한 검증을 통해 위의 결론들이 올바른 결론 임을 뒤받침 해준다.

6. 추가 과정을 통해 결정계수 값을 0.78에서 0.84까지 올렸다.
7. 추가 과정 뒤에 도출해 낸 회귀선에서는 쾌락추구, 중간 나이는 높을 수록 혁신 지수가 높았고, 불확실성에 대한 회피 지수는 낮을 수록 혁신 지수가 높았다.