FINAL ASSIGNMENT

Understanding Linear Regression and Its Implications

Instruction on the Final Assignment

최종 과제는 크게 다섯 개 파트로 나누어져 있습니다. 첫 번째 파트는 회귀분석의 가정에 대한 이해를 돕기 위한 것이며, 두 번째 파트는 회귀분석 모델에서 변수의 특성을 이해하기 위한 것입니다. 세 번째 파트는 회귀분석 모델에 있어서 변수 간 상관성에 대한 내용이며, 네 번째 파트는 시뮬레이션을 통한 경험적 접근에 관한 것입니다. 마지막으로 제시된 정치학 분야의 경험적 연구를 재현(replication)해보는 것입니다. 각각의 문제는 `RMarkdown` 또는 `Quarto`를 이용하여 PDF 형태로 제출하여 주시기 바랍니다. 혹은 PDF 컴파일에 성공하지 못하였다면, `.Rmd`나 `.qmd` 파일의 형태로 제출하셔도 좋습니다.

Data in Use Description

`V-Dem` Subset

변수는 모두 <민주주의의 다양성 (V-Dem)> 프로젝트의 버전 13 https://www.v-dem.net/en/data/data/v-dem-dataset/에서 추출한 것입니다.

Variable name	Description
country_name	국가명
country_id	국가별로 하나씩 주어지는 고유한 국가코드
year	관측 연도
v2x_polyarchy	이상적 선거민주주의를 얼마나 충족하고 있습니까?라는 질문에 대한 전문가 평가를 등간척도로 나타낸 변수. 선거민주주의지수(Electoral Democracy Index)라고도 함.
v2elembaut	선거관리기구(Election management body; EMB)가 국가 선거에 있어서 공정하게 선거법과 행정법률을 적용할 수 있는 자율성이 있습니까?라는 질문에 대한 전문가 평가를 등간척도로 나타낸 변수. 선거기구의 자율성(autonomy)이라고 볼 수 있다.
v2elembaut_ord	v2elembaut를 순서형 변수로 측정한 것
v2elembcap	선거관리기구(Election management body; EMB)가 국가 선거가 잘 운영될 수 있도록 충분한 직원과 자원을 가지고 있습니까?라는 질문에 대한 전문가 평가를 등간척도로 나타낸 변수. 선거기구의 역량(capacity)이라고 볼 수 있다.
v2elembcap_ord	v2elembcap을 순서형 변수로 측정한 것
v2elmulpar	국가 선거가 다당제 하에서 이루어지는지에 대한 전문가 평가를 등간척도로 나타낸 변수.

e_civil_war	각 국가-연도에 최소 1,000명 이상의 사상자가 발생한 국가 내 전쟁이 최소한 한 차례 있는 경우를 더미변수로 측정한 변수. Haber and Menaldo (2011)로부터 발췌.
e_pt_coup	Powell and Thyne (2011)에서 발췌한 변수로 (1) 실패한 쿠테타 (unsuccessful), (2) 성공한 쿠테타, (3) 쿠테타 시도의 세 값으로 측정됨. 쿠테타 시도는 군부 또는 국가기구 내의 다른 엘리트들에 의해 비헌법적 방법으로 현직 국가 수반을 하야시키고자 하는 전복 시도로 정의된다 (Powell and Thyne 2011, 252).
e_coups	주어진 연도 내에서 이루어진 성공적인 쿠테타의 수. 국가-연도 중에서 한 번 이상의 성공적 쿠테타의 관측이 이루어졌을 때, 그 해의 값은 최대값으로 측정된다.
e_gdppc	Maddison Project Database로부터 온 1인당 GDP 변수
e_pop	총 인구 수 (천명)
e_area	국토 면적 (제곱 킬로미터)
e_total_fuel	1인당 석유, 석탄, 그리고 천연가스 생산의 실질 가치, Haber and Menaldo (2011)로부터 발췌
e_regionpol_6C	지리적 지역을 보여주는 분류형 변수.
e_boix_regime	경쟁성(contestation)과 참여(participation)에 기초하여 Boix et al. (2013)가 측정한 이분형 민주주의 지수(0 or 1). (1) 정치 지도자가 자유롭고 공정한 선거로 선출되며, (2) 최소 수준의 보편선거권이 보장되는 국가일 경우 민주주의로 코딩되었음.
v2svindep	주어진 해에 그 국가가 독립국가인지 여부를 보여주는 더미변수.

```
## 데이터를 불러와 봅시다
library(devtools)
library(tidyverse)
if (!require(vdemdata)) install_github("vdeminstitute/vdemdata")
library(vdemdata)
vdem_sub <- vdemdata::vdem |>
  dplyr::select(country_name, country_id, year,
                v2x_polyarchy, v2elembaut, v2elembaut_ord,
                v2elembcap, v2elembcap_ord,
                v2elmulpar, e_civil_war,
                e_coups, e_pt_coup, e_pop, e_area,
                e_total_fuel = e_total_fuel_income_pc, e_regionpol_6C,
                e_boix_regime,
                e_gdppc, v2svindep)
```

Part I. Assumptions of OLS

문제 1

$\beta_0, \beta_1, \hat{\beta}_0, \hat{\beta}_1, x, x_i, u, \hat{u}_i, \bar{x}$는 각각 무엇을 의미하나요? 각 항(term)이 선형회귀모델에서 어떻게 해석되며, 각 항의 값은 무작위(random; 확률적)인지 또는 고정되어 있는 것(fixed)인지 설명하세요.

문제 2

문제 2-1 `lm()` 함수를 이용하여 로그값을 취한 1인당 GDP를 종속변수로, EMB capacity를 예측변수로 하는 회귀모델을 만들어보세요. `summary()` 함수를 이용해 모델을 요약하고 그 결과를 제시해보세요.
주의: 모델을 분석하기에 앞서 두 변수 모두에 대해 결측치(missing values)를 제거한 서브셋(subset)을 만들어 분석에 이용하세요.

문제 2-2 1인당 GDP의 원자료와 로그값을 취한 1인당 GDP가 EMB capacity와 갖는 관계를 산포도(scatter plot)를 통해 제시해주세요. 이때, 산포도는 회귀선을 포함하여 함께 보여주어야 합니다.

문제 3

아래의 통계치들을 R을 계산기로 활용하여 수기로(manually by hand) 계산하고 보고해주세요.

문제 3-1 모델의 제곱합(total sum of squares)

문제 3-2 모델의 잔차의 총 제곱합(total sum of squares)

문제 3-3 모델의 설명량의 총 제곱합(regression sum of squares)

문제 3-4 모델의 상관계수의 제곱(square of the correlation coefficient)

문제 3-5 모델의 조정된 R^2(adjusted R-square)

문제 3-6 EMB capacity의 계수값(coefficient)

문제 3-7 모델의 절편(intercept)

문제 4

잔차의 분포를 그래프로 제시하고 그것이 어떠한 의미를 가지는지 설명해주세요.

문제 5

표본 수준의 회귀함수(Sample regression function), $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$가 있다고 할 때, x는 $x = \log(z)$와 같이 또 다른 변수 z로부터 변형된 변수라고 하겠습니다(예를 들어 z는 연간 소득과 같은 큰 값을 지닌 변수일 수 있습니다).

문제 6-1 x가 한 단위 변화할 때, $\hat{y}$는 어떻게 변화하나요?

문제 6-2 z가 한 단위 변화할 때, $\hat{y}$는 어떻게 변화하나요?

x와 z의 변환(transformation) 관계가 회귀함수에 있어 의미하는 바가 무엇인지를 설명해 주세요.

Part II. Different Variables in OLS

문제 1

선거민주주의지수(electoral democracy index; EDI)를 종속변수로, 로그값을 취한 1인당 GDP(logged GDP per capita)를 예측변수로 하는 회귀모델을 추정하고 그 결과를 해석해 보세요.

문제 2

문제 1와 동일한 모델에 EMB autonomy 변수를 추가하여 추정해보세요.

문제 2-1. 결과를 해석해보세요.

문제 2-2. 모델 1과 모델 2의 차이를 서술해보세요.

문제 3

아래와 같은 절차를 거쳐 EMB autonomy에 대한 부분기울기계수(partial slope coefficient)를 찾아보세요.

1. EDI를 종속변수로, 로그값을 취한 1인당 GDP를 예측변수로 하는 회귀모델을 분석
2. EMB autonomy를 종속변수로, 로그값을 취한 1인당 GDP를 예측변수로 하는 회귀모델을 분석
3. 1단계에서 분석한 모델의 잔차를 종속변수로, 2단계에서 분석한 모델의 잔차를 예측변수로 하는 회귀모델을 분석
4. 1단계와 2단계로부터 얻은 각 잔차 간의 관계를 그래프로 제시
5. 1-3단계까지의 절차가 무엇을 의미하는지 간단하게 설명

문제 4

문제 2와 동일한 모델에 EMB capacity 변수를 추가하여 추정해보세요.

문제 4-1. 결과를 해석해보세요.

문제 4-2. 문제 2의 모델과 문제 4의 모델의 차이를 서술해보세요
(효과의 크기와 R^2를 반드시 언급할 것)

문제 5

```
if (!require(DAAG)) install.packages("DAAG")
library(DAAG)
```

{DAAG} 패키지의 vif() 함수를 이용하여 다음의 분석을 수행하세요.

1. 각각의 예측변수에 대해 다중공선성(multicollinearity)을 확인하고 분산팽창인자(Variance inflation factors, VIFs)를 보고
2. EMB capacity를 종속변수로 하고 나머지 공변인(GDP와 EMB autonomy)을 예측변수로 하는 회귀분석을 수행
3. 앞의 2단계에서 수행한 분석으로부터 R^2를 계산하고 그것을 이용해 EMB capacity에 대한 VIF를 계산
4. GDP를 종속변수로, 나머지 공변인(EMB autonomy and capacity)을 예측변수로 하는 회귀분석을 수행
5. 앞의 4단계에서 수행한 분석으로부터 R^2를 계산하고 그것을 이용해 GDP에 대한 VIF를 계산

6. 3단계와 5단계를 바탕으로 VIF 값의 차이가 의미하는 바에 대해 설명

문제 6

내전 발발 여부(e_civil_war)에 관한 더미변수를 포함하여 문제 1에서와 같은 모델을 추정해보세요. 그 결과를 해석하고 본인이 생각할 때, 내전 발발 여부와 선거민주주의지수 간의 관계를 가장 잘 보여줄 것이라고 기대되는 그래프를 제시해보세요

문제 7

지역(factor(e_regionpol_6C)라는 함수를 이용하여)에 대한 분류형 변수를 추가하여 문제 1에서와 같은 모델을 추정하고 그 결과를 해석해보세요. 본인이 생각할 때, 지역과 선거민주주의지수 간의 관계를 가장 잘 보여줄 것이라고 기대되는 그래프를 제시해보세요.

Part III. Relationship among Variables in OLS

문제 1

Brambor, Clark, and Golder (2006) 를 읽고, 저자들의 주장을 자세히 설명해보세요. 본인이 분석하고 싶은 데이터를 하나 선택하여 연속형 변수와 연속형 변수의 상호작용항을 하나, 연속형 변수와 이산형 변수의 상호작용항을 포함한 모델을 각각 하나씩 특정(specify)한 뒤, Brambor, Clark, and Golder (2006) 의 조언에 따라 분석을 수행해보세요. 유의미한 그래프를 그리고 그에 대해 설명해보세요.

문제 2

문제 2의 경우 Berry, Burden, and Howell (2010) 의 논문을 읽고 논문의 데이터를 다운받고, 아래의 모든 분석을 전통적인 방법으로, 매우 자세하게 논리적으로 설명해보세요.

Berry, Burden, and Howell (2010) 데이터에 대한 정보

> 각 관측치는 지역구-연도의 분석수준을 가진다. high_lnoutlays_cpi는 주어진 지역구-연도에 있어서 시민 한 명 당 배정되는 연방 비용에 자연로그를 취한 값이다. party 변수는 하원의원의 정당이 어디인지를 나타낸다. 미국이기 때문에 1이면 공화당(Republican), 0이면 민주당(Democrat)이라고 하자. year는 관측치의 연도를 의미한다. president는 해당 의원이 현재 집권정당 소속인지 여부를 나타내는 더미변수이다. leader는 해당 의원이 다수당 대표인지(leader), 하원의장인지(speaker), 원내 총무인지(whip)를 나타내는 명목형 변수이다. party_distance 는 하원의원의 자기이념과 소속정당의 평균이념의 거리를 측정한 변수이다. freshman은 초선의원인지 여부를 나타내는 변수이며, any_chair 와 any_rank 는 해당 의원이 어떤 위원회든지 간에 위원장 또는 간사를 맡고 있는지 여부를 나타내는 변수이다. party_size 는

정당의 규모를 의미하며, 마지막으로 여러 mem_ 변수들은 해당 의원이 속한 위원회를 보여준다.

문제 2-1. 논문의 데이터는 STATA 파일의 확장자인 .dta의 포맷을 가지고 있습니다. 따라서 적절한 방법을 통해 해당 파일을 R의 객체로 로드해보세요.

문제 2-2. high_lnoutlays_cpi, 연방지출(federal spending)을 종속변수로 하는 모델을 하나만 특정해보세요. 그 통계모델에는 president, party_distance 와 그 두 변수의 상호작용항, party_size , 그리고 이외에 중요하다고 생각되는, 또는 종속변수를 설명하기 위해 필요한 다른 변수들을 포함하세요. 각각의 변수에 대한 회귀계수가 무엇을 의미하는지 서술하고, R이 제공하는 모델에 관한 통계치들을 해석하세요.

문제 2-3. 국회 농림축산식품해양수산위원회에 박 의원이라는 가상의 인물이 있다고 합시다. 만약 그가 농림축산식품해양수산위원회에서 예산결산특별위원회로 자리를 옮기고자 한다면 지역구에 가져갈 수 있는 연방 지출의 규모(take-home)는 어떻게 변화할까요? 그 변화는 실질적으로 또는 통계적으로 유의미한 변화일까요?

문제 2-4. 박 의원이 점점 더 보수적으로 변하고 있다고 합시다. 그는 점점 자신의 정당으로부터 이념적으로 멀어지고 있습니다(party_distance). 그의 이념이 정당 이념으로부터 점점 더 멀어지는 이 현상이 그가 연방 예산으로부터 지역구로 가져가는 예산(take-home)에 어떠한 영향을 미칠까요? 또한 박 의원이 집권당 소속인지의 여부(president)에 따라 지역구에 가져가는 예산 규모의 변화가 있는지 분석해보세요.

문제 3

Berry, Burden, and Howell (2010) 의 데이터를 이용하여 high_lnoutlays_cpi를 종속변수로 하고 문제 2에서 사용한 동일한 변수들에 party_size의 제곱항과 party_size와 president의 상호작용항을 추가하여 회귀분석을 수행해보세요.

문제 3-1. 모델의 결과를 서술하세요.

문제 3-2. party_size가 미국 하원(U.S. House)의 구성원에게 있어서 지역구에 가져가는 예산 규모(take-home)에 미치는 영향에 대해 전반적인 분석을 수행하라. 통계적 유의성에 대한 검정을 실시하세요.

Part IV. Simulations

문제 1

자유롭게 부트스트래핑(bootstrapping)이 무엇인지 검색하여 찾아보고 이해한 바에 대해 정리하여 서술해보세요.

Non-parametric Bootstrapping

문제 2

문제 2-1 변수 EMB capacity를 결측치가 존재하지 않도록 전처리하고 이를 모집단(population)이라고 가정합시다. 이때, EMB capacity 변수의 분포를 그래프로 표현하고 총 관측치의 수와 평균이 얼마인지를 제시해보세요.

문제 2-2 복원추출(random sampling with replacement)로 무작위 표집을 하여 다음과 같은 결과를 제시하세요.

문제 2-2-1 모집단 EMB capacity로부터 10개씩의 관측치를 갖는 표본 10개를 추출하고 그들 각각의 평균을 객체에 저장하세요.

문제 2-2-2 모집단 EMB capacity로부터 10개씩의 관측치를 갖는 표본 1,000개를 추출하고 그들 각각의 평균을 객체에 저장하세요.

문제 2-2-3 모집단 EMB capacity로부터 1,000개씩의 관측치를 갖는 표본 10개를 추출하고 그들 각각의 평균을 객체에 저장하세요.

문제 2-2-4 모집단 EMB capacity로부터 1,000개씩의 관측치를 갖는 표본 1,000개를 추출하고 그들 각각의 평균을 객체에 저장하세요.

위의 네 표집에 대해 각각의 표본평균들의 분포를 그래프로 나타내고 그 기대값(expected value)가 얼마인지를 제시해보세요. 위의 네 방법 중 어떤 것이 모집단의 모수–EMB capacity에 보다 정확한 결과와 적은 불확실성(작은 분산)을 보이는지 순위를 매겨보세요.

문제 2-3 EMB capacity를 종속변수로 하고 자연로그값을 취한 1인당 GDP, 민주주의, 내전 발발, 그리고 EMB capacity를 예측변수로 하는 선형회귀모델을 추정하고 그 결과를 해석하세요.

문제 2-4 복원추출을 이용한 무작위 표집으로 기존 데이터셋에서 10,000개의 관측치를 가지는 표본을 추출한 다음 위의 문제 2-3의 모델을 재추정하세요. 추정 결과가 얼마나 달라졌는지 설명하세요.
(단, 복원추출을 이용한 무작위 표집을 하기 이전에 모델에 투입되는 변수들이 결측치를 갖지 않게 하는 서브셋을 만들고 시작하세요. 그렇지 않으면 표집된 10,000개의 관측치 결과에도 결측치들이 포함될 수 있습니다.)

문제 2-5 문제 2-4에서의 과정을 1,000번 반복한 이후에 예측변수 중 자연로그를 취한 1인당 GDP에 대한 계수값의 결과를 저장하세요. 제대로 시행하였다면 1,000개의 계수값을 가지게 될 것입니다. 그 결과를 그래프로 나타내고, 계수들의 분포를 서술하세요.
(이때, 1,000번 반복하는 과정은 시간이 좀 걸릴 수 있습니다.)

Parametric Bootstrapping

문제 3

King, Tomz, and Wittenberg (2000) 를 읽고, 논문에서 저자들이 주장하고, 제안하는 바에 대해 정리해보세요. 구체적으로 1) 제목이 의미하는 바가 무엇인지, 2) 예측값(predicted values)과 기대값(expected values)의 차이가 무엇인지, 3) 베이지언 접근법에 대한 그들의 설명에 대해 논하라. 이에 대한 답은 구글을 포함한 어떤 출처든 자유롭게 사용하여 서술하세요.

문제 4

문제 4-1 다변량 정규확률분포(multivariate normal distribution)가 무엇인지 자유롭게 검색하여 자신이 이해한 바를 서술하세요.

문제 4-2 다음의 R 코드를 따라 실행해보고 그 결과가 의미하는 바가 무엇인지를 서술하세요. 다변량 정규확률분포에 관한 이해와 연관지어서 서술하세요.

```
link <- "https://stats101.netlify.app/assignments/STATS101Dataset.csv"
library(ezpickr)
data <- ezpickr::pick(link)

model4 <- lm(v2elembcap ~ log(e_migdppc) +
               v2x_polyarchy + e_civil_war + v2elembaut,
             data = data)

library(mvtnorm)
beta_draws <- rmvnorm(4000, mean = coef(model4),
                      sigma = vcov(model4))
```

```r
head(beta_draws)
```

```
     (Intercept) log(e_migdppc) v2x_polyarchy e_civil_war v2elembaut
[1,]   -2.743731      0.3371835     0.5102926  -0.4874688  0.5035630
[2,]   -2.809767      0.3521523     0.4154217  -0.3886043  0.5046395
[3,]   -2.806744      0.3469558     0.4607599  -0.3944397  0.5038402
[4,]   -2.853826      0.3533836     0.4885990  -0.3719250  0.5026109
[5,]   -2.715022      0.3448793     0.3334204  -0.4258327  0.5306773
[6,]   -2.855833      0.3541529     0.4283872  -0.3730365  0.5025056
```

```r
apply(beta_draws, 2, mean)
```

```
   (Intercept) log(e_migdppc)  v2x_polyarchy     e_civil_war      v2elembaut
    -2.7796749      0.3465082      0.4358110     -0.4221434      0.5106516
```

```r
apply(beta_draws, 2, quantile, probs = c(0.025, 0.975))
```

```
      (Intercept) log(e_migdppc) v2x_polyarchy e_civil_war v2elembaut
2.5%    -2.950424      0.3237611     0.2940238  -0.5007363  0.4888638
97.5%   -2.605125      0.3692177     0.5790280  -0.3449466  0.5322667
```

문제 4-2-1 beta_draws는 무엇인가요?

문제 4-2-2 beta_draws의 평균과 2.5, 97.5분위 값은 각각 무엇을 의미하나요?

문제 4-2-3 beta_draws에 포함된 각 변수들의 분포를 그래프로 표현하세요.

Part V. Understanding OLS in Practice via Replications

문제 1

상당히 공을 들인 통계 분석 결과를 마주했다고 합시다. 데이터의 크기도 상당하고, 가우스-마르코프와 전통적 선형모델에 대한 가정들도 모두 잘 들어맞는다고 할 때, 이론적으로 관심을 가지고 있는 변수에 대한 p-값이 꽤 큰, $p = .70$ 정도의 결과를 확인하였다고 하겠습니다. 엄밀하게 통계적으로 이 결과에 대해서 어떻게 결론을 내릴 수 있을까요? 실질적으로 이 결과에 대해서 말할 수 있는 두 가지 경우는 무엇이 있을까요(왜 p-값이 크게 나타났을까요)? p-값이 크게 나타날 수 있는 두 가지 경우를 한 번에 보여줄 수 있는 가장 단순한 방법이 무엇이 있을까요?

문제 2

Bell, Clay, and Martinez Machain (2017) 가 *Journal of Conflict Resolution*에 개제한 논문, "The Effect of US Troop Deployments on Human Rights"를 일독하고 이 논문의 재현 데이터(replication files)를 찾아서 논문의 Table 1에 해당하는 분석과 그래프를 재현해보세요.

- 우선 구글에서 Bell, Clay, and Martinez Machain (2017) 의 데이터를 찾아 어떻게 R에서 열 수 있는지를 찾아보세요
- Bell, Clay, and Martinez Machain (2017) 의 연구는 우리가 아직 배우지 않은 고급 통계기법을 사용하고 있기 때문에, 우리가 할 수 있는 방법을 이용하여 재현해보도록 합시다.
- Bell, Clay, and Martinez Machain (2017) 의 코드는 STATA 코드로 이루어져 있기 때문에 이를 R로 재현하는 과정에는 일종의 "번역"이 필요합니다.
- 배운 기법을 이용해 어떻게 그들의 연구를 재현할 것인지 설명하세요.
- 비슷한 결과를 재현해내었나요?
- 재현한 결과의 효과의 크기가 논문의 그래프와 유사한가요?
- 모수적 부트스트랩(parametric bootstrap)을 이용하여 불확실성을 보여주고, 그 결과를 설명해보세요.

References

Bell, Sam R., K. Chad Clay, and Carla Martinez Machain. 2017. "The Effect of US Troop Deployments on Human Rights." *Journal of Conflict Resolution* 61 (10): 2020–42. https://doi.org/10.1177/0022002716632300.

Berry, Christopher R., Barry C. Burden, and William G. Howell. 2010. "The President and the Distribution of Federal Spending." *American Political Science Review* 104 (4): 783–99.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63–82. https://doi.org/10.1093/pan/mpi014.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 341355.