

데이터와 사회과학통계

2020 여름 국민대학교 행정학과

0. 데이터와 사회과학

0. 데이터와 사회과학

1. 빅데이터 시대와 사회과학

(1) 지난 세기의 사회과학

0. 데이터와 사회과학

1. 빅데이터 시대와 사회과학

(1) 지난 세기의 사회과학

(2) 빅데이터 하에서 달라지는 것

- 자료의 규모
- 자료의 구성
- 자료의 분석 준거

0. 데이터와 사회과학

1. 빅데이터 시대와 사회과학

(1) 지난 세기의 사회과학

(2) 빅데이터 하에서 달라지는 것

(3) 사회과학의 새로운 패러다임?

0. 데이터와 사회과학

1. 빅데이터 시대와 사회과학

(1) 지난 세기의 사회과학

(2) 빅데이터 하에서 달라지는 것

(3) 사회과학의 새로운 패러다임?

2. A house is (built) with stones

• 앙리 프앙카레

“돌을 쌓아 집을 만드는 것처럼,
과학은 자료로 만들어진다. 그러나
돌무더기가 집이 아니듯 자료
의 더미가 과학은 아니다”

• 더 읽어보기

한신갑. (2015). 빅데이터와 사회과학하기: 자료기반의 변화와 분석전략의 재구상. *한국사회학*, 49(2), 161-192.

0. 데이터와 사회과학

1. 빅데이터 시대와 사회과학

(1) 지난 세기의 사회과학

(2) 빅데이터 하에서 달라지는 것

(3) 사회과학의 새로운 패러다임?

2. A house is (built) with stones

• 목표

- 사회과학 연구의 기초
가설검증의 논리와 통계적 기법
단순회귀
다중회귀
로지스틱 회귀

• R을 통한 데이터 분석

- 실제 데이터를 가지고 다중회귀/
로지스틱 회귀모형을 통해 분석
 - 이론적 근거가 있어야 하고, 이에 대한 짧은 발표

I. 사회과학 연구의 기초

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

1) 명제와 가설

- “교육수준과 행복 사이에는 어떤 관계가 있을까?”

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

1) 명제와 가설

- 개념화(conceptualization)와 조작화(operationalization)

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

1) 명제와 가설

2) 가설의 검증

- 연구 대상(target population)을 설정

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

1) 명제와 가설

2) 가설의 검증

- 인과성 causality

- Aristotle

"We think we have scientific knowledge when we know the cause."

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

1) 명제와 가설

2) 가설의 검증

• 어떻게 통제하는가?

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

- 1) 명제와 가설
- 2) 가설의 검증
- 3) 변수의 유형

- 모형에서의 역할에 따른 구분
 - 독립변수 independent variable
 - 종속변수 dependent variable
 - 통제변수 control variable

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

- 1) 명제와 가설
- 2) 가설의 검증
- 3) 변수의 유형

• 변수의 성격에 따른 구분

- 연속변수 continuous variable
- 범주형변수 categorical variable

I. 사회과학 연구의 기초

1. 사회과학 연구의 기초

(1) 사회과학연구 과정에 대한 이해

1) 명제와 가설

2) 가설의 검증

3) 변수의 유형

• 변수의 결합과 분석방법

변수의 결합 형태		분석방법
이분/다분-이분/다분		Chi-square
이분-연속		t-test
다분-연속		분산분석 ANOVA
연속-연속		상관분석 correlation
종속변수	연속	선형회귀분석
	이분	이항로지스틱 회귀분석
	다분	다항로지스틱 회귀분석

II. R Basics

II. R Basics

1. R이란?

II. R Basics

1. R이란?

2. Packages

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# Set your working directory  
setwd("C:/PAPP")
```

```
# Verify that your working directory is set  
correctly  
getwd()
```

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# Create a numeric and a character variable
```

```
a <- 5
```

```
typeof(a)
```

```
a
```

```
b <- "kookmin university"
```

```
typeof(b)
```

```
b
```

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# Create a vector
```

```
my.vector <- c(10,-7,99,34,0,-5)
```

```
my.vector
```

```
length(my.vector)
```

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# Subsetting vector
```

```
my.vector[1]
```

```
my.vector[-1]
```

```
my.vector[2:4]
```

```
my.vector[c(2,5)]
```

```
my.vector[length(my.vector)]
```

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# Deleting
```

```
rm(a)
```

```
rm(list=ls())
```


II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# create a matrix
```

```
my.matrix1 <- matrix(data =  
c(1,2,30,40,500,600), nrow = 3, ncol = 2,  
byrow = TRUE, dimnames = NULL)
```

```
my.matrix2 <- matrix(data =  
c(1,2,30,40,500,600), nrow = 2, ncol = 3,  
byrow = FALSE)
```

```
my.matrix1
```

```
my.matrix2
```

```
# subsetting a matrix
```

```
my.matrix1[1,2]
```

```
my.matrix1[2,1]
```

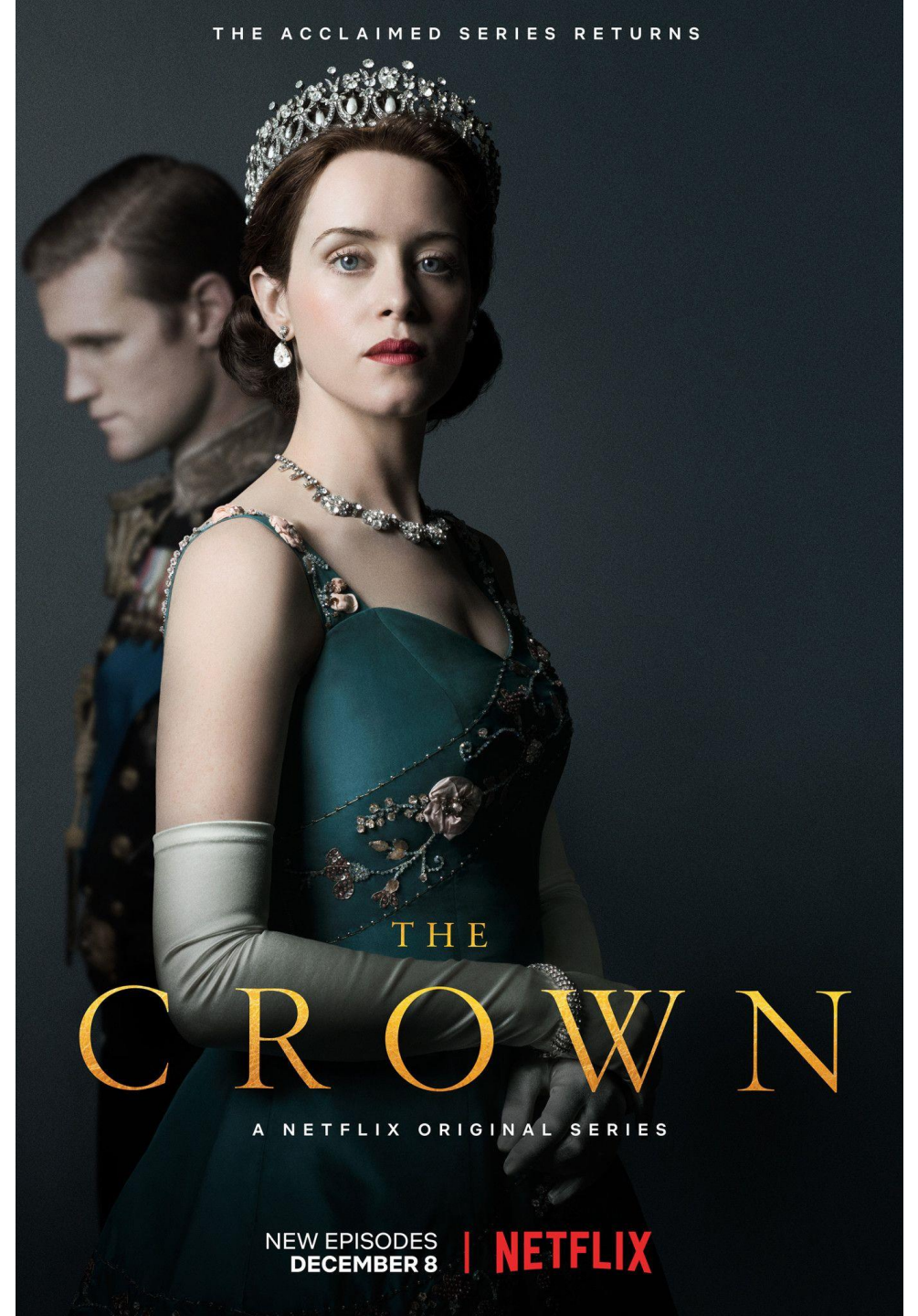
```
my.matrix1[,1]
```

```
my.matrix1[1:2,]
```

```
my.matrix1[c(1,3),]
```

II. R Basics

1. R이란?
2. Packages
3. R syntax



II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# Polity IV dataset
```

```
my.data <- read.csv("polity.csv")
```

```
dim(my.data)
```

```
my.data[1:10,]
```

```
names(my.data)
```

```
levels(my.data$country)
```

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
# drop
```

```
my.data <- my.data[my.data$year==1946,]
```

```
my.data[1:10,]
```

```
summary(my.data$polity2)
```

```
table(my.data$nato, my.data$polity2)
```

```
summary(my.data$polity2[my.data$nato==0]) #  
not in nato
```

```
summary(my.data$polity2[my.data$nato==1]) #  
nato member
```

II. R Basics

1. R이란?

2. Packages

3. R syntax

```
## illustration
```

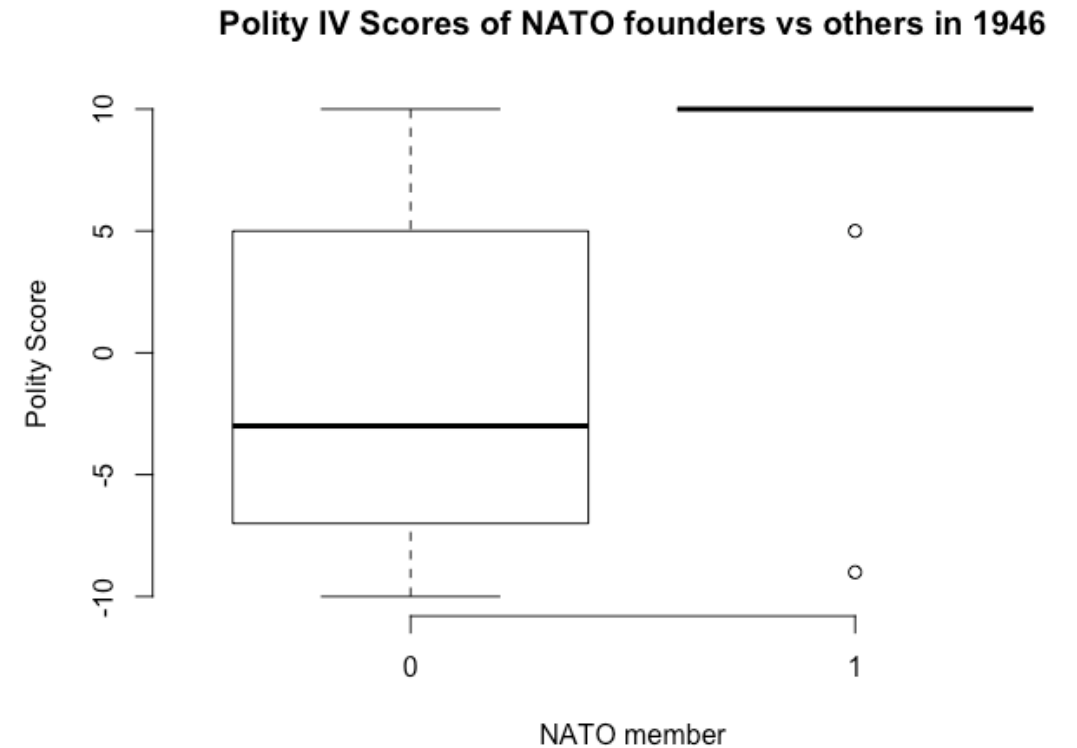
```
boxplot(my.data$polity2 ~  
as.factor(my.data$nato), frame = FALSE, main =  
"Polity IV Scores of NATO founders vs others  
in 1946", xlab = "NATO member", ylab =  
"Polity Score")
```

II. R Basics

1. R이란?

2. Packages

3. R syntax



II. R Basics

1. R이란?

2. Packages

3. R syntax

```
install.packages(c("maps", "mapdata"))
```

```
library(maps)
```

```
library(mapdata)
```

```
map(database = 'world', region = c('South  
Korea', 'North Korea'))
```

```
points (126.996845,  
37.612262 ,col=2,pch=20,cex=1.8)
```

III. 가설검증의 논리와 통계적 기법

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

- 동전맞추기 게임과 귀류법

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

- 귀무가설과 대립가설
 - H_0 = 이 동전은 보통 동전이다.
 - H_1 = 이 동전은 보통 동전이 아니다. 즉 특수 동전이다.

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

- 유의수준과 p-value

- $p - \text{value} \leq \alpha$: 귀무가설 기각
- $p - \text{value} > \alpha$: 귀무가설 기각 못함

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

- 중심극한정리

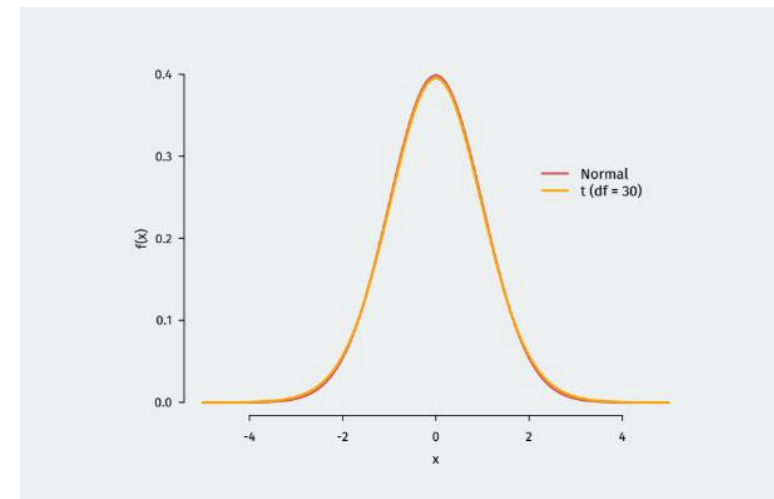
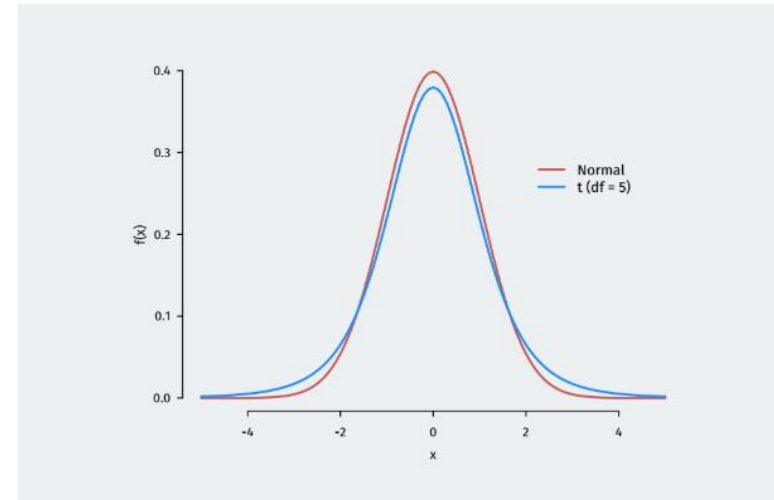
III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

• T-분포와 자유도



III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

$$\bullet t = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

- 귀무가설: 두집단 간의 평균에 차이가 없다

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

- “공무원의 연령대에 따라 직무만족도가 다른가?”

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

• 분산의 의미

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

- 어떻게 분산을 분석해서 평균 차이를 확인할 수 있는가?

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

• 2개의 상황

- 상황1: 평균이 동일한 세 개의 집단이 있음. 세 집단에서 각각 n 개씩 무작위로 뽑아서 세 개의 표본집단을 만들었다.
- 상황2: 평균이 다른 세 개의 집단이 있음. 세 집단에서 각각 n 개씩 무작위로 뽑아서 세 개의 표본 집단을 만들었다.

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

$$\bullet F = \frac{\text{집단간분산 (between groups)}}{\text{집단내분산 (within groups)}}$$

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

- $F = \frac{\text{집단간분산 (between groups)}}{\text{집단내분산 (within groups)}}$
- 집단내분산이 크면
 - 평균들의 차이에 대한 자신감 떨어짐 (반비례)
 - 설명 안된 분산
- 집단간분산이 크면
 - 평균들의 차이에 대한 자신감 커짐 (비례)
 - 설명된 분산

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

- ANOVA의 대립가설과 귀무가설
 - 귀무가설: 20대=30대=40대 (상황 1)
 - 대립가설: 20대 \neq 30대 혹은 30대 \neq 40대 혹은 40대 \neq 20대 (상황2)

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

- 상관관계분석의 목적

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

• 어떻게 측정?

- scatter plot

- correlation coefficient, r

$$\bullet r = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

- 상관계수의 의미

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

• 공분산의 의미

$$• S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

• 상관관계의 해석

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# Set your working directory  
setwd("C:/PAPP")
```

```
# Verify that your working directory  
is set correctly  
getwd()
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
#  
rm(list=ls())  
  
# foreign file formats  
library(foreign)  
world.data <- read.dta("QoG2012.dta")  
  
# the dimensions  
dim(world.data)  
  
# the variable names  
names(world.data)  
  
#  
head(world.data)
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# dplyr package  
install.packages("dplyr")  
library(dplyr)
```

```
# rename h_j to judiciary  
world.data <- rename(world.data,  
  judiciary = h_j)
```

```
names(world.data)
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# frequency table  
table(world.data$judiciary)
```

```
# creating a factor variable  
world.data$judiciary <-  
factor(world.data$judiciary,  
labels = c("independent",  
"controlled"),  
levels = c(1, -5))
```

```
#  
head(world.data)
```

```
#  
table(world.data$judiciary)
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
#
summary(world.data$wdi_gdpc)

# creating subsets
free.legal <- filter(world.data,
judiciary == "independent")

controlled.legal <-
filter(world.data, judiciary ==
"controlled")

# remove missings
mean(free.legal$wdi_gdpc, na.rm =
TRUE)

mean(controlled.legal$wdi_gdpc,
na.rm = TRUE)
```


III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# t.test  
t.test(world.data$wdi_gdpc ~  
world.data$judiciary, mu=0,  
alt="two.sided", conf=0.95)
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

welch Two Sample t-test

```
data: world.data$wdi_gdpc by world.data$judiciary
t = 6.0094, df = 98.261, p-value = 3.165e-08
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 7998.36 15885.06
sample estimates:
mean in group independent mean in group controlled
      17826.591             5884.882
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# renaming  
world.data <- rename(world.data, hdi  
= undp_hdi)
```

```
world.data <- rename(world.data,  
corruption.control = wbgi_cce)
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# scatterplot
plot(x =
world.data$corruption.control,
      y = world.data$hdi,
      xlim = c(xmin = -2, xmax = 3),
      ylim = c(ymin = 0, ymax = 1),
      frame = FALSE,
      xlab = "World Bank Control of
Corruption Index",
      ylab = "UNDP Human Development
Index",
      main = "Relationship b/w
Quality of Institutions and Quality
of Life")
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r



III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

```
# Pearson's r  
cor.test  
(world.data$corruption.control,  
world.data$hdi,  
use="complete.obs",  
conf.level = 0.99)
```

III. 가설검증의 논리와 통계적 기법

0. hypothesis testing

1. t-test

(1) t 분포

(2) t-test의 논리

2. 분산분석 ANOVA

(1) ANOVA의 목적

(2) ANOVA의 논리

(3) F-test

3. 상관관계분석 correlation

(1) 상관계수, r

Pearson's product-moment correlation

```
data: world.data$corruption.control and  
world.data$hdi
```

```
t = 12.269, df = 173, p-value < 2.2e-16
```

```
alternative hypothesis: true correlation is not equal  
to 0
```

```
99 percent confidence interval:
```

```
0.5626121 0.7736905
```

```
sample estimates:
```

```
cor
```

```
0.6821114
```

IV. 단순회귀모형

IV. 단순회귀 모형

0. 회귀분석의 이해

- 회귀분석의 정의

IV. 단순회귀 모형

0. 회귀분석의 이해

- 회귀분석의 기능

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

- 독립변수가 1단위 증가했을 때, 종속변수는 얼마나 변할까?
- $y = a + bx$

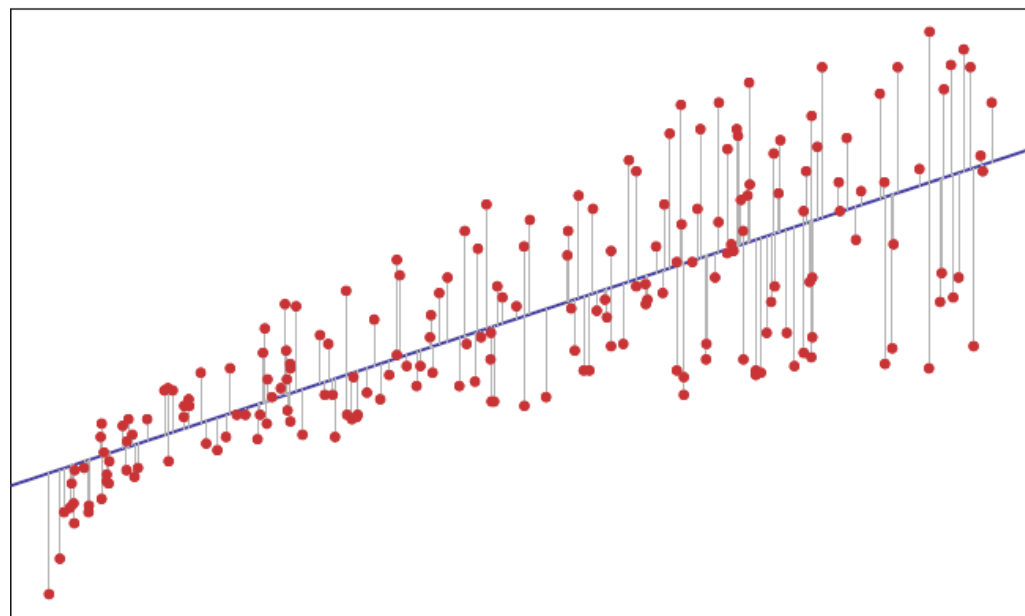
IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

- 오차항과 오차제곱합
- 오차 $e_i =$ 관측치 y_i - 예측치 \hat{y}_i



- $y = a + bx + e$

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

- 오차제곱합 sum of squares error SSE
- Minimize $\sum e_i^2 =$
Minimize $\sum (y_i - \hat{y}_i)^2$

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

- 기울기 (b)

- $$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

- 절편 (a)

- $\bar{y} = a + b\bar{x} \rightarrow a = \bar{y} - b\bar{x}$

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

$$\bullet \text{ 월급} = 125.893 + 3.014 \times \text{연령}$$

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

- 상관계수와 회귀계수 간의 관계

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

- 상관계수-회귀계수에 대한 오해

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

3. 통계적 유의도

(1) 단순회귀모형에서의 결정계수

- “주어진 자료”를 가장 잘 설명 \neq 독립변수가 종속변수를 매우 잘 설명

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

3. 통계적 유의도

(1) 단순회귀모형에서의 결정계수

• 결정계수(r^2)의 의미

- 회귀선에 기초한 오차제곱합
- 종속변수의 평균에 기초한 오차제곱합

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

3. 통계적 유의도

(1) 단순회귀모형에서의 결정계수

- 왜 둘을 비교하나?

- $r^2 =$

$$1 - \frac{\text{회귀식에 기초한 오차제곱합(설명안된 분산)}}{\text{종속변수의 평균에 기초한 오차제곱합(총분산)}}$$

$$(0 \leq r^2 \leq 1)$$

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

3. 통계적 유의도

(1) 단순회귀모형에서의 결정계수

(2) 유의도 검증

- 표본에서 구한 회귀식을 모집단에 적용하려면?
- 종속변수의 분산 = 설명된 분산 (sum of squares of regression: SSR = 회귀제곱합)
- + 설명안된 분산 (sum of squares error: SSE = 오차제곱합)
- = sum of squares of total: SST 총분산

IV. 단순회귀 모형

0. 회귀분석의 이해

1. 단순회귀모형의 이해

(1) 선형성과 회귀식의 추정원리

(2) 최소제곱법 ordinary least square

(3) 단순회귀모형의 해석

2. 상관분석과 단순회귀분석

3. 통계적 유의도

(1) 단순회귀모형에서의 결정계수

(2) 유의도 검증

- F-test

- $F_{(k, n-k-1)}$

$$= \frac{\text{회귀제곱합}(SSR)/k}{\text{오차제곱합}(SSE)/n - k - 1}$$

(n : 사례수, k : 독립변수의 수)

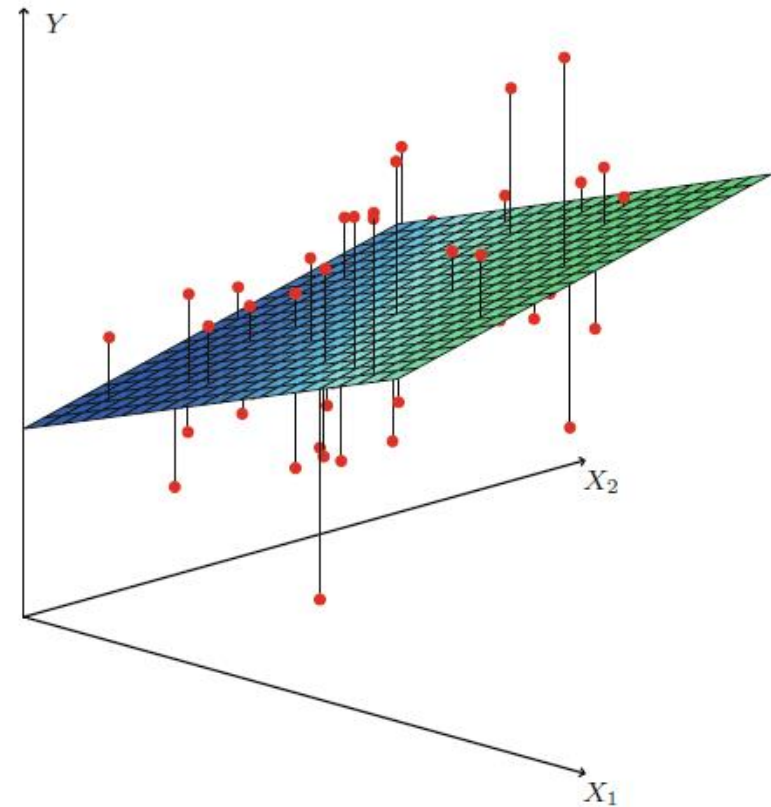
V. 다중회귀모형

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

$$\bullet y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n + e$$



V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

- 통계적 통제의 의미

- $$\text{월급여} = -241 + 4 * \text{연령} + 23$$
* 교육연수

V. 다중회귀모형

1. 다중회귀모형의 이해

- (1) 다중회귀모형의 기본형태
- (2) 다중회귀모형의 해석
- (3) 표준화 회귀계수

• 변수간의 영향을 비교하려면?

- $$\frac{\text{월급여} = -241 + 4 * \text{연령} + 23}{* \text{교육연수}}$$

V. 다중회귀모형

1. 다중회귀모형의 이해

- (1) 다중회귀모형의 기본형태
- (2) 다중회귀모형의 해석
- (3) 표준화 회귀계수

• Then how?

• $\beta = b \frac{s_y}{s_x}$

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

$$\bullet r^2 = 1 -$$

$$\frac{\text{회귀식에 기초한 오차제곱합}(SSE)}{\text{종속변수의 평균에 기초한 오차제곱합(총분산 } SST)}$$

$$= \frac{\text{총분산 } SST - \text{오차제곱합 } SSE}{\text{총분산 } SST} = \frac{\text{회귀분산 } SSR}{\text{총분산 } SST}$$

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

- Adjusted r^2

- $r^2 - \frac{k(1-r^2)}{n-k-1}$

(n : 표본수, k : 독립변수의 수)

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

- 모집단에 대한 적용 가능성

- $F_{(k, n-k-1)}$

$$= \frac{\text{회귀제곱합}(SSR)/k}{\text{오차제곱합}(SSE)/n - k - 1}$$

$$= \frac{\text{회귀제곱평균 } MSR}{\text{오차제곱평균 } MSE}$$

(n : 사례수, k : 독립변수의 수)

V. 다중회귀모형

1. 다중회귀모형의 이해

- (1) 다중회귀모형의 기본형태
- (2) 다중회귀모형의 해석
- (3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

- (1) 다중회귀모형에서의 결정계수
- (2) 다중회귀모형의 유의도 검증

- F-test

- 모든 독립변수들이 유의미하지 않다 vs. 독립변수 중 하나 이상은 유의미하다

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

• t-test

- 어느 독립변수가 유의미한 독립변수인가?

- $$\text{월급여} = -241 + 4 * \text{연령} + 23 * \text{교육연수}$$

- 두가지 오류 error를 고려해야 함
 - 표본추출오류 sampling error
 - 통제안된 분산 uncontrolled variation

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

- 신뢰구간을 통한 검증
confidence interval

- 개별 회귀계수의 표준오차
standard error

- $$SE = \sqrt{\frac{\text{오차제곱합 } SSE / (n-2)}{\sum (x_i - \bar{x})^2}}$$

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

- 가설검정을 통한 검증
hypothesis testing

- $$t = \frac{\text{회귀계수}(b) - 0}{SE} = \frac{\text{회귀계수}(b)}{SE}$$

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

- 어떤 변수를 추가시켜 전체모형의 설명력이 증가하면, 증가된 설명력은 곧 추가된 변수의 설명력이다?
 - 통제안된 분산 uncontrolled variation 존재
 - 결정계수를 구하는 과정

V. 다중회귀모형

1. 다중회귀모형의 이해

- (1) 다중회귀모형의 기본형태
- (2) 다중회귀모형의 해석
- (3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

- (1) 다중회귀모형에서의 결정계수
- (2) 다중회귀모형의 유의도 검증
- (3) 다중회귀분석에 대한 오해들

- t-test, ANOVA, 단순회귀모형에서 유의미하게 나온 독립변수만 선택하여 다중회귀모형을 구성하면 된다?

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

- t-test, ANOVA, 단순회귀모형에서 유의미하게 나온 독립변수만 선택하여 다중회귀모형을 구성하면 된다?

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
# Set your working directory
```

```
setwd("C:/PAPP")
```

```
# Verify that your working directory is set  
correctly
```

```
getwd()
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
#  
install.packages("remotes")  
library(remotes)  
install_github("cran/zeig")  
install.packages("texreg")  
  
library(texreg)  
library(zeig)  
library(dplyr)
```


V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
#
```

```
rm(list = ls())
```

```
#
```

```
world_data <-  
read.csv("qog_std_cs_jan15.csv")
```

```
#
```

```
dim(world_data)
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
#  
world_data <- select(world_data,  
  country = cname,  
  political_stability = wbgpse,  
  latitude = lp_lat_abst,  
  globalization = dr_ig,  
  inst_quality = ti_cpi)
```

```
#  
head(world_data)
```

```
#  
summary(world_data)
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

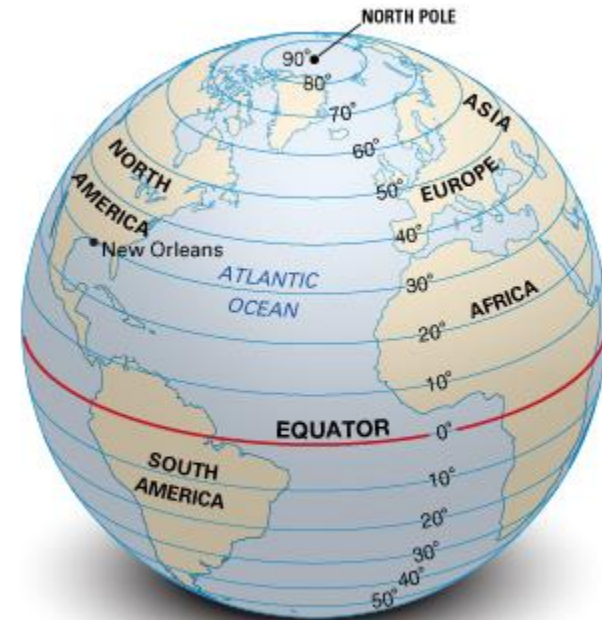
3. 다중회귀분석 소결

```
#
```

```
world_data <- filter(world_data,  
                        !is.na(latitude),  
                        !is.na(globalization),  
                        !is.na(inst_quality))
```

```
#
```

```
summary(world_data)
```



V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
# transform
```

```
world_data$latitude <- world_data$latitude *  
90
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
#
```

```
plot(political_stability ~ latitude, data =  
world_data)
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

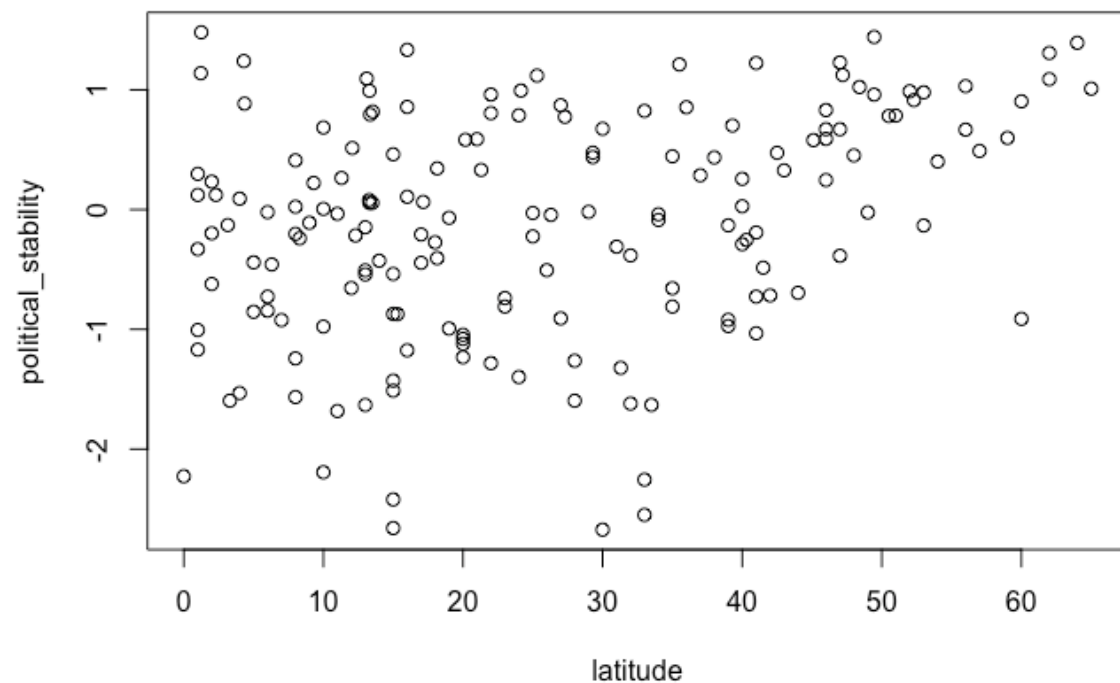
2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결



V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
#
```

```
latitude_model <- lm(political_stability ~  
latitude, data = world_data)
```

```
#
```

```
plot(political_stability ~ latitude, data =  
world_data)
```

```
#
```

```
abline(latitude_model)
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

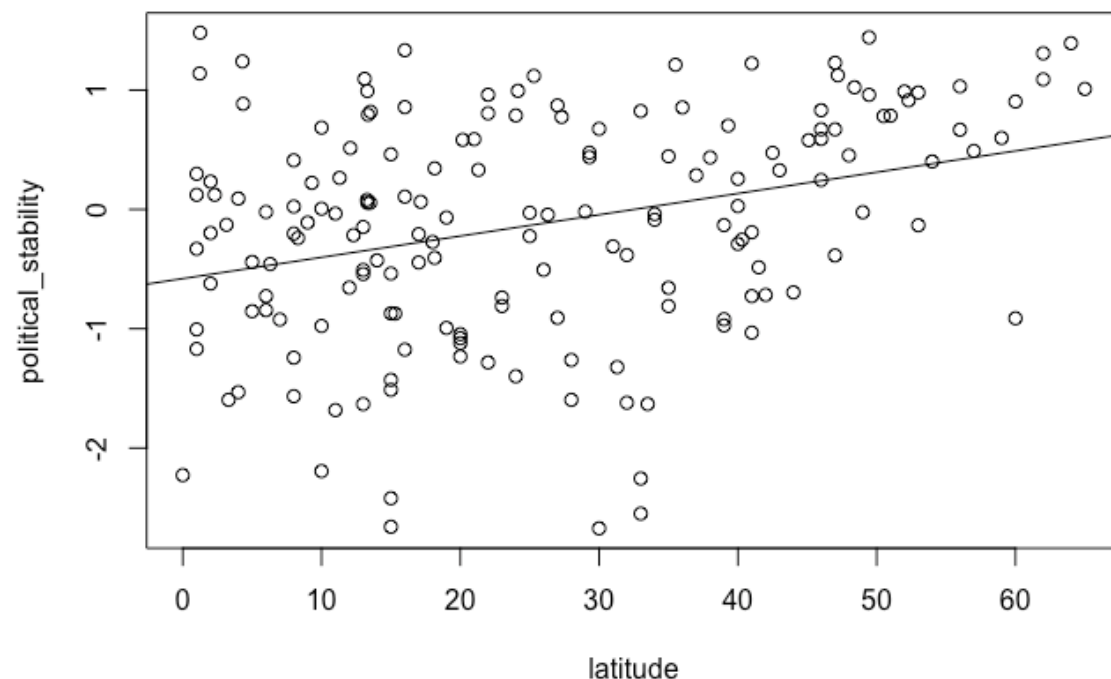
2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결



V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
# regression output  
screenreg(latitude_model)
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
=====
                        Model 1
-----
(Intercept)    -0.58 ***
                  (0.12)
latitude        0.02 ***
                  (0.00)
-----
R^2              0.11
Adj. R^2         0.10
Num. obs.       170
RMSE             0.89
=====
```

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
# new model
```

```
inst_model <- lm(political_stability ~  
latitude + globalization + inst_quality,  
data = world_data)
```

```
#
```

```
screenreg(list(latitude_model, inst_model)))
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

	Model 1	Model 2
(Intercept)	-0.58 *** (0.12)	-1.25 *** (0.20)
latitude	0.02 *** (0.00)	0.00 (0.00)
globalization		-0.00 (0.01)
inst_quality		0.34 *** (0.04)
R^2	0.11	0.50
Adj. R^2	0.10	0.49
Num. obs.	170	170
RMSE	0.89	0.67

*** p < 0.001, ** p < 0.01, * p < 0.05

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
#
```

```
anova(latitude_model, inst_model)
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

Analysis of Variance Table

Model 1: political_stability ~ latitude

Model 2: political_stability ~ latitude + globalization +
inst_quality

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	168	133.12				
2	166	74.28	2	58.841	65.749	< 2.2e-16 ***

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

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

```
# multivariate regression model
```

```
z.out <- zelig(political_stability ~  
latitude + globalization + inst_quality,  
data = world_data, model = "ls")
```

```
# setting covariates
```

```
x.out <- setx(z.out, inst_quality = seq(0,  
10, 1))
```

```
# simulation
```

```
s.out <- sim(z.out, x = x.out)
```

```
# plot results
```

```
ci.plot(s.out, ci = 95, xlab = "Quality of  
Institutions")
```

V. 다중회귀모형

1. 다중회귀모형의 이해

(1) 다중회귀모형의 기본형태

(2) 다중회귀모형의 해석

(3) 표준화 회귀계수

2. 다중회귀모형의 통계적 유의도

(1) 다중회귀모형에서의 결정계수

(2) 다중회귀모형의 유의도 검증

(3) 다중회귀분석에 대한 오해들

3. 다중회귀분석 소결

