

8장 회귀분석

U S I N G



STATISTICAL ANALYSIS FOR
SOCIAL SCIENCE USING R

1. 회귀분석의 적용
2. 단순회귀분석의 분석 방법
3. 다중회귀분석의 분석 방법

1. 회귀분석의 적용

1. 변수들의 척도

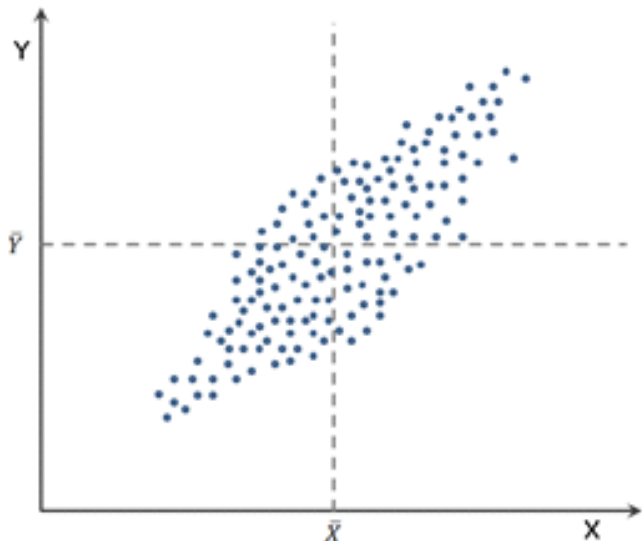
- 독립변수와 종속변수 간의 인과관계를 검증하기 위한 방법
 - ✓ 독립변수와 종속변수는 **모두 등간척도나 비율척도로** 측정된 변수
 - ✓ 종속변수의 수는 1개
 - 독립변수의 수가 **1개인** 회귀분석은 단순회귀분석(Simple regression)
 - 독립변수의 수가 **2개 이상인** 회귀분석은 다중회귀분석(Multiple regression)

2. 회귀분석의 통계량

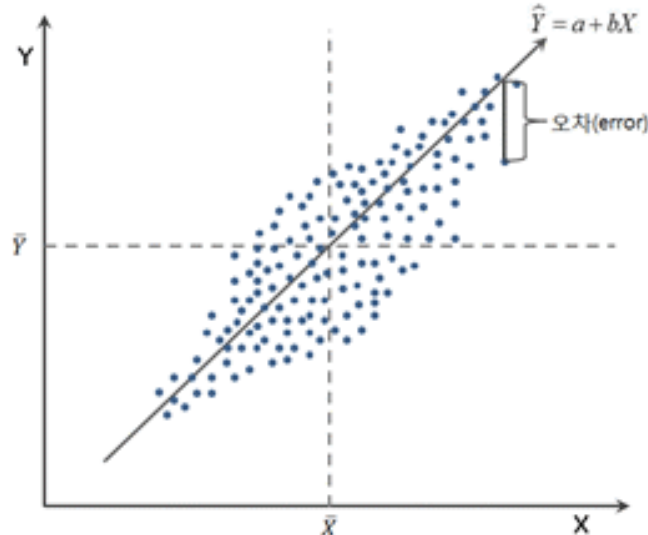
- 독립변수 X 와 종속변수 Y 간의 관계
 - ✓ X 는 독립변수의 값, \hat{Y} 는 독립변수가 특정 값일 경우에 **예상되는** 종속변수 Y 의 값(a 는 절편, b 는 기울기)

$$\hat{Y} = a + bX$$

1. 회귀분석의 적용



(a) 산포도



(b) 일차함수로 표현한 변수의 관계

- 일차 함수로 추정된 종속변수의 값과 실제 자료 간에는 차이가 나타남
 - ✓ 실제 자료를 일차함수의 형태로 표현하게 되면 오차(e)를 포함하게 됨

$$Y = a + bX + e$$

1. 회귀분석의 적용

- 독립변수와 종속변수의 관계를 가장 잘 나타낼 수 있는 일차함수
 - ✓ 최소자승법(Ordinary Least Square)을 통해 오차가 가장 최소화될 수 있는 일차함수를 찾아냄

$$\sum_{i=1}^N (e_i)^2 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N \{Y_i - (a + bx_i)\}^2$$

- ✓ a, b 로 각각 편미분하여 통해 절편(a)과 기울기(b)를 구함

$$b = \frac{Cov(X, Y)}{Var(X)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

$$\begin{aligned} \sum e_i^2 &= \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 X_i)^2 \\ \frac{dQ}{db_0} &= 2 \sum (Y_i - b_0 - b_1 X_i) (-1) = 0 \Rightarrow \sum Y_i - n b_0 - b_1 \sum X_i = 0 \Rightarrow b_0 = \frac{1}{n} \sum Y_i - b_1 \frac{1}{n} \sum X_i \\ \frac{dQ}{db_1} &= 2 \sum (Y_i - b_0 - b_1 X_i) (-X_i) = 0 \\ &= \sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0 \Rightarrow \sum X_i Y_i - (\bar{Y} - b_1 \bar{X}) \sum X_i - b_1 \sum X_i^2 = 0 \\ &\Rightarrow \sum X_i Y_i - \bar{Y} \sum X_i = b_1 (\sum X_i^2 - \bar{X} \sum X_i) \\ &\Rightarrow b_1 = \frac{\sum X_i (Y_i - \bar{Y})}{\sum X_i^2 - \bar{X} \sum X_i} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum [(X_i - \bar{X})^2]} \\ &\Rightarrow \sum X_i (Y_i - \bar{Y}) \Rightarrow \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \bar{Y} \sum X_i - \bar{X} \sum Y_i + n \bar{X} \bar{Y} \\ &= \sum X_i Y_i - \bar{Y} n \bar{X} - \bar{X} n \bar{Y} + n \bar{X} \bar{Y} \\ &= \sum X_i Y_i - \sum X_i \bar{Y} \\ &= \sum X_i (Y_i - \bar{Y}) \end{aligned}$$

Handwritten notes and corrections:

- $\sum X_i - n \bar{X} = 0$
- $\sum (X_i - \bar{X}) = 0$
- $\sum (X_i - \bar{X})(X_i - \bar{X}) = \sum (X_i - \bar{X})^2$
- $\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \bar{Y} \sum X_i - \bar{X} \sum Y_i + n \bar{X} \bar{Y}$

1. 회귀분석의 적용

- 회귀계수(기울기)의 유의성 검증
 - ✓ 회귀계수의 **추정치**를 **표준오차**로 나눈 값이 자유도 $n-(k+1)$ 을 갖는 t 분포를 따름 (n 은 자료의 개수, k 는 독립변수의 개수)

Significance test \rightarrow

$$t_{(n-k-1)} = \frac{b_i}{S_e(b_i)} = \frac{b_i}{\frac{S_e}{\sqrt{\sum (X_i - \bar{X})^2}}}$$

Standard Error of b_i

Standard Error of Estimation \rightarrow

$$S_e = \sqrt{\frac{SSE}{n - k - 1}}$$

degree of freedom \rightarrow

$$df = n - k - 1$$

1. 회귀분석의 적용

- 결정계수(R^2)
 - ✓ 종속변수의 전체 제곱합(SST)은 독립변수에 의해 설명되는 제곱합(SSR; 회귀제곱합)과 독립변수에 의해 설명되지 않는 제곱합(SSE; 오차제곱합)으로 구성
 - 회귀제곱합이 크다는 것은 회귀계수가 크다는 것이고, 독립변수가 종속변수에 큰 영향을 미친다는 것을 의미

$$\sum (Y - \bar{Y})^2 = \sum (Y - \hat{Y} + \hat{Y} - \bar{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2$$

$$SST = SSE + SSR$$

- ✓ R^2 는 종속변수의 전체 변동 중에서 독립변수에 의해 설명된 변동의 비율로, 회귀모델의 설명력을 나타냄

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

1. 회귀분석의 적용

- 회귀모델의 전체적인 유의성 검증
 - ✓ 분산분석과 동일하게 F 분포를 활용

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F ratio
Regression	SSR	$(k+1)-1 = k$	$MSR = SSR/k$	MSR/MSE
Error	SSE	$n-(k+1) = n-k-1$	$MSE = SSE/(n-k-1)$	
Total	SST	$n-1$	$MST=SST/(n-1)$	



1. 회귀분석의 적용

- 다중회귀분석

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

- ✓ 독립변수들 간의 관계는 **독립적인 관계를** 가정

- 절편과 회귀계수를 추정할 때 독립변수들 간에는 전혀 관계가 없는 것을 가정하고 계산
- 하나의 독립변수가 종속변수에 미치는 영향력을 의미하는 회귀계수는 나머지 다른 독립변수를 모두 통제된 후에 계산된 회귀계수

- ✓ 다중공선성(Multicollinearity)의 문제

- 독립변수들 간에 강한 상관관계가 있다면 독립변수들 간에 서로 관계가 있는 분산은 제외하고 나머지 독립변수의 분산으로 회귀계수를 구해야 하기 때문에 잘못된 회귀계수가 구해질 수 있음
- 다중공선성의 문제는 독립변수들 간의 상관계수를 구하거나 분산팽창요인(Variance Inflation Factor)을 구하여 진단

$$VIF_i = \frac{1}{1 - R_i^2}$$

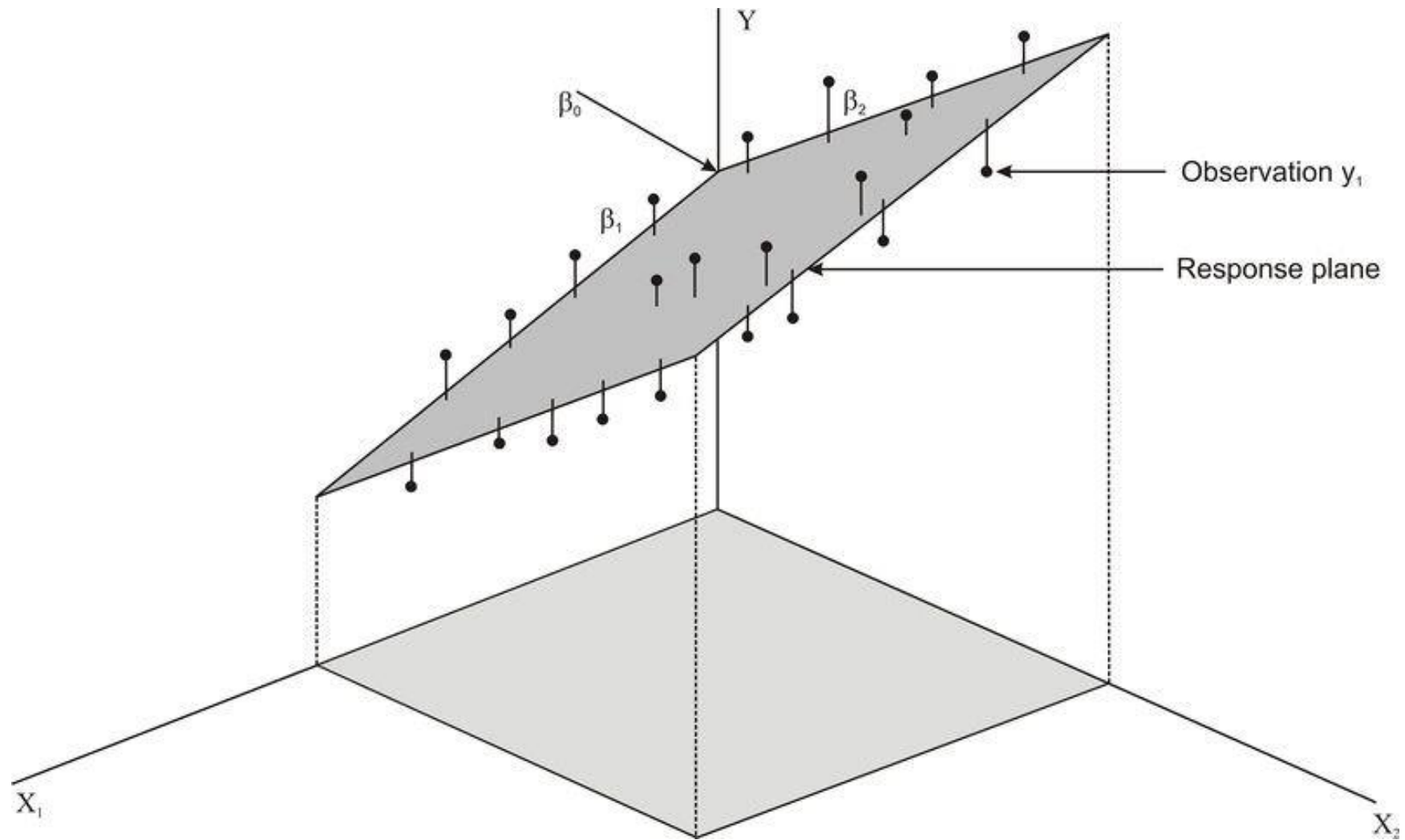
» 독립변수들 간의 강한 상관계수가 나타나거나 분산팽창요인의 값이 10을 넘는다면, 상관관계가 높은 독립변수 중 **일부를 제거하거나**, 높은 상관관계의 독립변수를 **하나의 변수로 변형하는** 등의 방법으로 해결

1. 회귀분석의 적용

- ✓ 2개 이상의 독립변수들 중에서 종속변수에 가장 큰 영향을 미치는 영향 비교
 - 독립변수의 측정 단위가 다르기 때문에 회귀계수만으로 직접 비교할 수 없음
 - 대신 **표준화된 회귀계수(β)**를 통해 독립변수가 종속변수에 영향을 미치는 정도를 비교
- 더미변수(dummy variables)
 - ✓ 독립변수가 범주형 변수일 경우에 더미변수로 변환하여 분석 가능
 - ✓ 더미변수는 집단의 특성에 따라 0과 1로 변환하여 만들어지고, 만들 수 있는 더미변수의 갯수는 (집단의 수-1)
 - ✓ 더미변수를 이용한 회귀분석에서는 준거집단을 어떤 집단으로 가정할 것인가에 따라서 결과가 달라지고, **준거집단을 기준으로** 결과를 해석해야 하기 때문에 준거집단이 매우 중요

	더미변수1(D1)	더미변수2(D2)	
변수값=(1)	D1=0	D2=0	준거집단
변수값=(2)	D1=1	D2=0	
변수값=(3)	D1=0	D2=1	

I. 회귀분석의 적용







.권용재

1일 전

read.csv 로 데이터 프레임을 읽어들이때 컬럼 이름 깨짐 현상.

read.csv 로 데이터를 불러들일때 첫 컬럼 이름이 깨지는 현상을 혹시 colnames 같이 재지정하는 방법 말고 다른방법 아시는 분 계신가요..?

1.PNG (1.073 Kb)

댓글

인용

수정

삭제

이메일

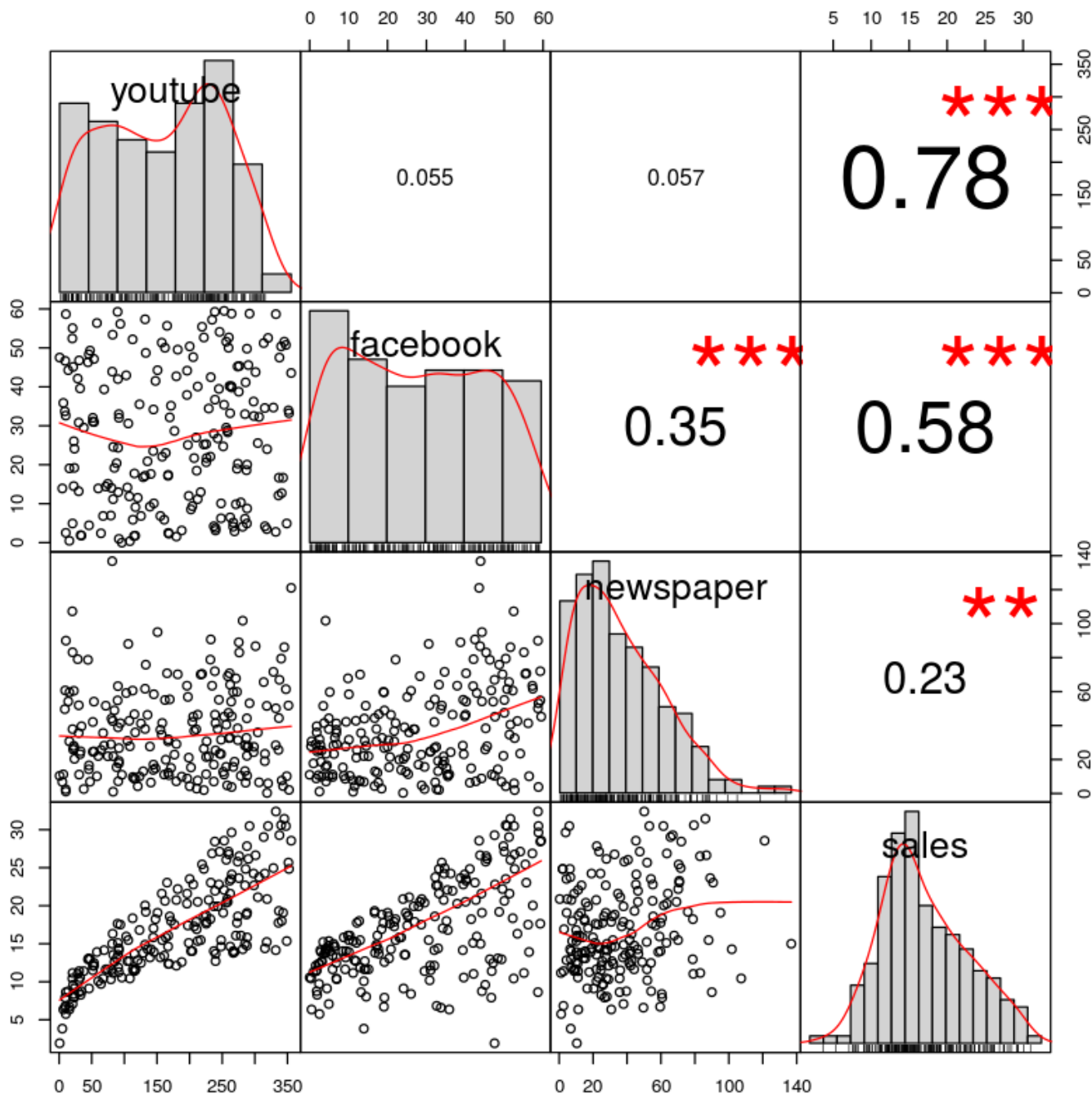
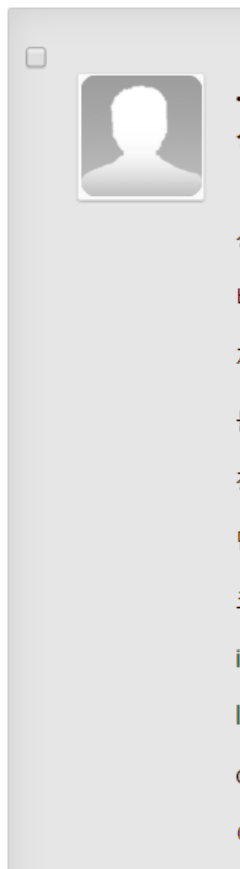


.한상혁


23시간 전

댓글: read.csv 로 데이터 프레임을 읽어들이때 컬럼 이름 깨짐 현상.

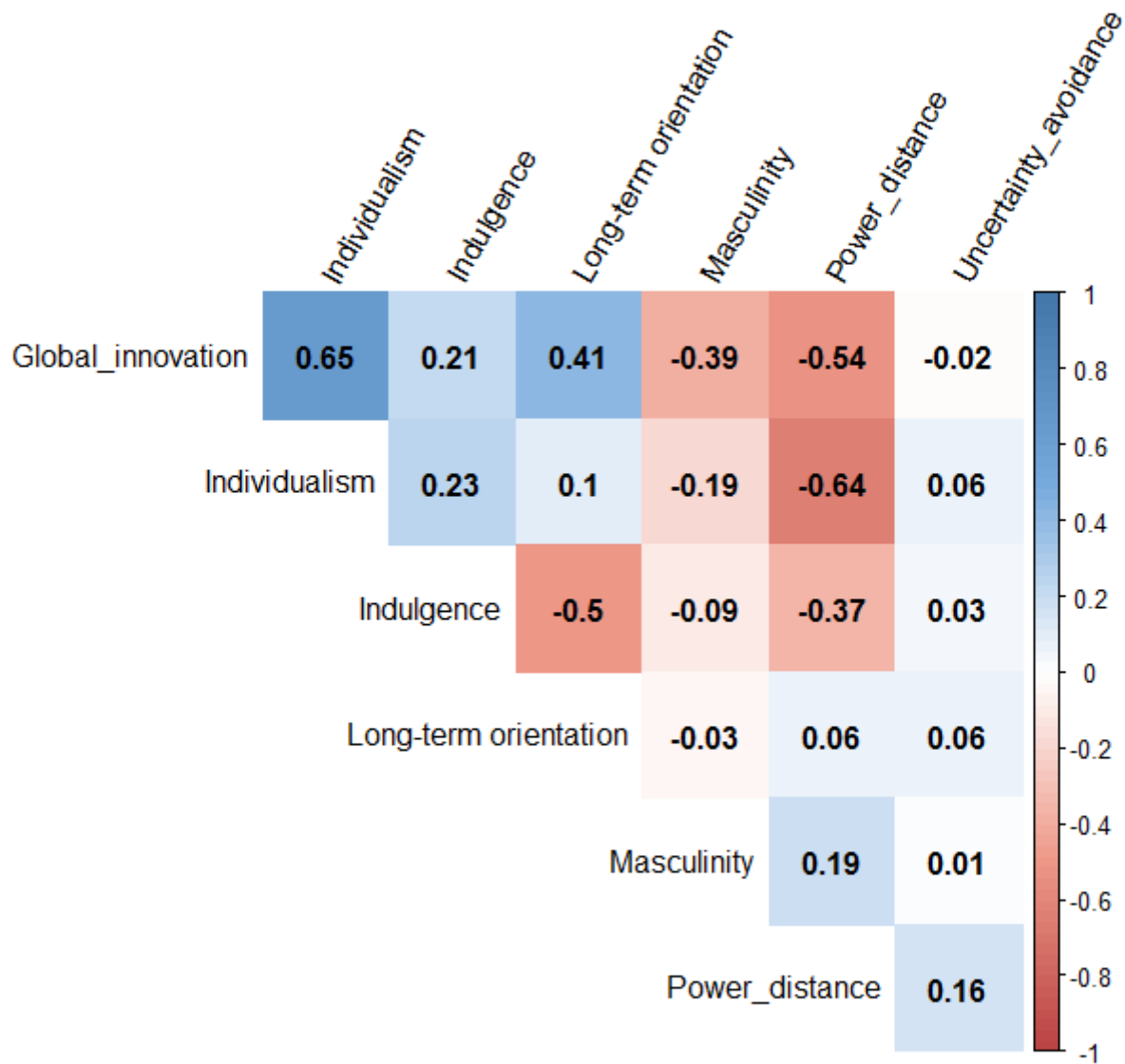
인코딩 문제라면 read.csv에서 인코딩 관련 인자들을 조정하는 방법이 있긴 한데 정확히 무슨 문제인지, 인코딩이 무엇인지 파악하려면 저 데이터가 있어야 할 거 같아요.



9시간 전

☐


.한
 싱
 if(
 in
 lib
 }
 결
 co
 > c
 [1]
 co
 co
 = '
 만



23시간 전

50, tl.cex=1, insig



2. 단순회귀분석의 분석 방법

연구가설

- 1) 자기신뢰감은 자아존중감에 영향을 미칠 것이다.
- 2) 부모에 대한 애착은 자아존중감에 영향을 미칠 것이다.

1. 연구가설 1의 검증

<분석 순서>

- 1) 연구가설의 검증을 위해서 사용할 변수는 이미 만들어진 상태에 있기 때문에 그 변수들을 사용한다.
- 2) 자기신뢰감이 자아존중감에 미치는 영향에 대한 단순회귀분석을 수행한다.
- 3) 'sjPlot' 패키지를 이용해서 회귀분석 결과표와 도표를 출력한다.

```
# ② 단순회귀분석
# 연구가설 1의 검증
regression1_1 <- lm(self.esteem ~ self.confidence, data=spssdata)
summary(regression1_1)
# 별도의 객체로 저장하지 않는 분석 방법 : 위의 명령문과 같은 결과가 산출됨
summary(lm(self.esteem ~ self.confidence, data=spssdata))
```




2. 단순회귀분석의 분석 방법

- ✓ 회귀분석을 시행하기 위해 lm 함수를 이용한다.
 - **lm 함수에는** 종속변수와 독립변수를 차례대로 입력하고, 종속변수와 독립변수 사이에는 '~' 표시
 - 연구가설 1은 종속변수가 자아존중감이고, 독립변수가 자기신뢰감 이므로 self.esteem ~ self.confidence라고 입력하고, 종속변수와 독립변수가 있는 데이터를 지정하면 된다(data=spssdata). 그리고 lm 함수에서 지정한 **회귀모형은 regression1_1이라는 객체에 할당**
 - summary 함수에 이 객체를 지정하면 회귀분석 결과를 확인
 - 회귀분석 결과를 객체로 저장하지 않고 분석결과를 보기 위해서 summary 함수와 lm 함수를 함께 사용하여 회귀모형의 결과를 확인

2. 단순회귀분석의 분석 방법

```
> regression1_1 <- lm(self.esteem ~ self.confidence,
data=spssdata)
> summary(regression1_1)
```

Call:
lm(formula = self.esteem ~ self.confidence, data = spssdata)

Residuals:

	Min	1Q	Median	3Q	Max
	-12.3028	-2.1606	-0.1285	2.1651	12.8073

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.08255	0.72102	18.145	<2e-16 ***
self.confidence	0.58716	0.06608	8.886	<2e-16 ***

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

Residual standard error: 3.451 on 593 degrees of freedom
Multiple R-squared: 0.1175, Adjusted R-squared: 0.116
F-statistic: 78.95 on 1 and 593 DF, p-value: < 2.2e-16

*Adjusted R-squared

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2



2. 단순회귀분석의 분석 방법

- 분석결과
 - ✓ 절편의 추정량(Intercept)은 13.08255이고, 독립변수의 기울기 값에 대한 추정량은 0.58716
 - ✓ 독립변수의 회귀 계수에 대한 통계적 검증을 위한 통계량은 **T 값**
 - T 값은 8.886이고, T 값에 대한 유의도는 $2e-16$ 로 소수점 15번째 자리까지 0인 값
 - 유의도가 영가설을 채택 혹은 기각할 수 있는 기준인 0.05보다 낮은 수준이므로 영가설은 기각되고, 연구가설을 채택
 - ✓ R^2 값은 0.1175로 종속변수의 전체 분산 중에서 독립변수에 의해 설명되는 분산이 11.75%
 - ✓ **F 값**은 78.950이고, 유의도는 $2.2e-16$
 - F 값에 따른 유의도가 영가설의 기각 혹은 채택의 기준인 0.05보다 낮은 수준이므로 영가설은 기각



2. 단순회귀분석의 분석 방법

```
# ③ 'sjPlot' 패키지를 이용한 회귀분석 결과표와 도표 출력
library(sjPlot)
tab_model(regression1_1, show.se = T, show.ci = F, pred.labels =
c("(Intercept)", "자기신뢰감"), dv.labels = c("자아존중감"))
tab_model(regression1_1, show.se = T, show.ci = F)
tab_model(regression1_1, show.se = T, show.ci = F, pred.labels =
c("(Intercept)", "자기신뢰감"), dv.labels = c("자아존중감"), file =
"simple_regression.html")
file.show("simple_regression.html")
```

- ✓ 연구가설 1-1에 대한 회귀분석 결과를 'sjPlot' 패키지의 tab_model 함수를 이용하여 출력
 - tab_model 함수에서는 앞서 lm 함수로 회귀분석을 한 결과를 할당한 객체 regression1_1을 이용하여 결과표로 출력
 - 결과표에 출력될 통계량을 지정
 - pred.labels에는 독립변수의 label, dv.labels에는 종속변수의 label을 지정
 - 출력될 결과표를 외부 파일로 저장하기 위해 file 인자로 파일을 저장할 경로와 파일 이름을 입력

2. 단순회귀분석의 분석 방법

자아존중감			
<i>Predictors</i>	<i>Estimates std. Error</i>		<i>p</i>
(Intercept)	13.08	0.72	<0.001
자기신뢰감	0.59	0.07	<0.001
Observations	595		
R ² / adjusted R ²	0.117 / 0.116		

- 분석결과
 - ✓ 앞서 lm 함수를 통해 자기신뢰감이 자아존중감에 미치는 영향을 살펴본 결과와 동일한 결과가 출력

*회귀분석의 기본 가정

1. 정규성(Normality)

- 종속변수의 관측치가 **주어진** 독립변수들의 선형결합에 대해서 정규분포를 따름 (For each X_i)

2. 등분산성(Equality of Variances)

- 종속변수의 관측치가 **모든** 독립변수들의 선형결합에 **걸쳐서** 동일한 분산을 가짐 (For all X_i)

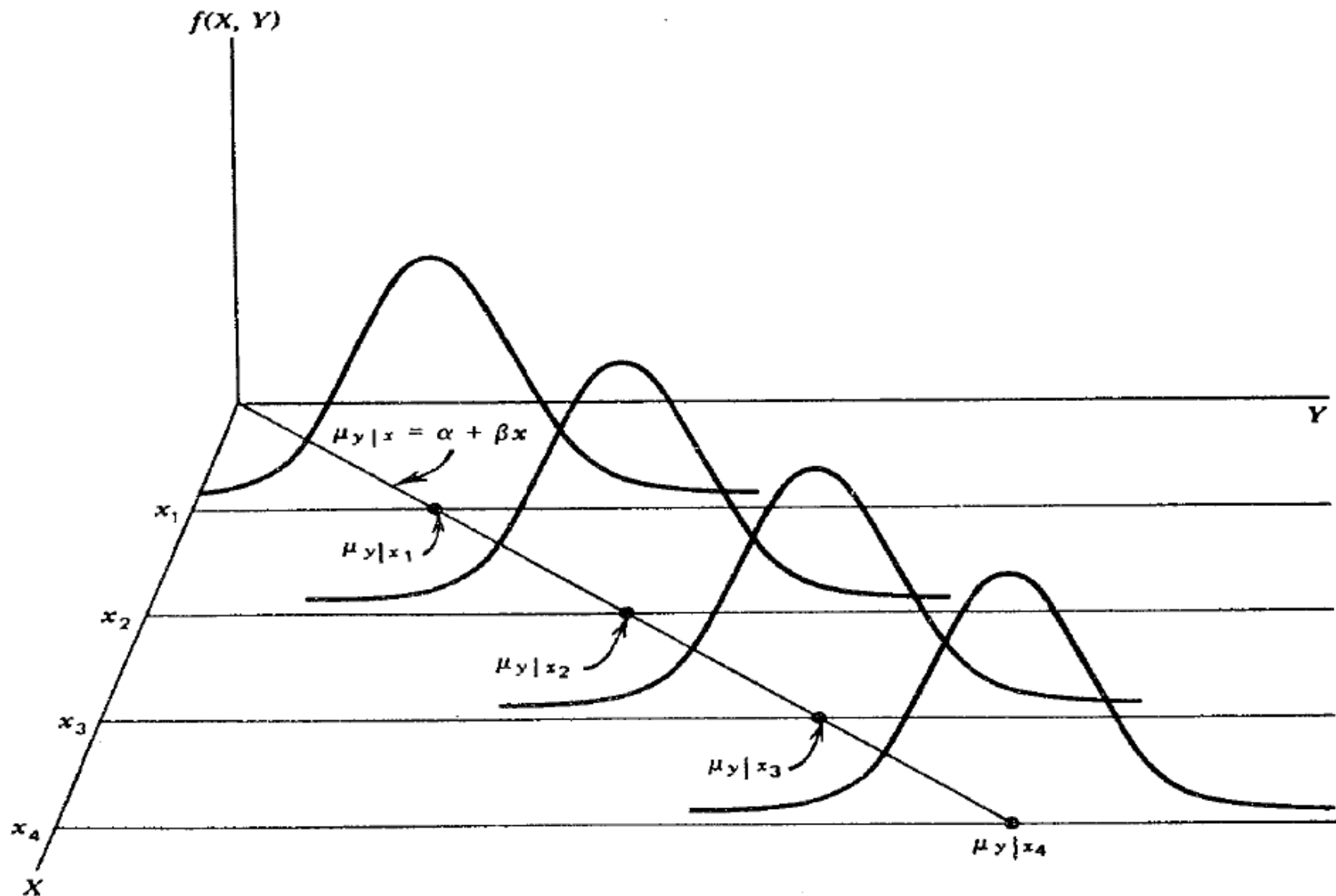
3. 독립성(Independence; no autocorrelation)

- 서로 다른 독립변수들의 선형결합에 대한 종속변수의 관측치는 상호 독립적임 (For all X_i, X_j ; i not equal j)

4. 선형성(Linearity)

- 독립변수와 종속변수 간의 관계는 선형임

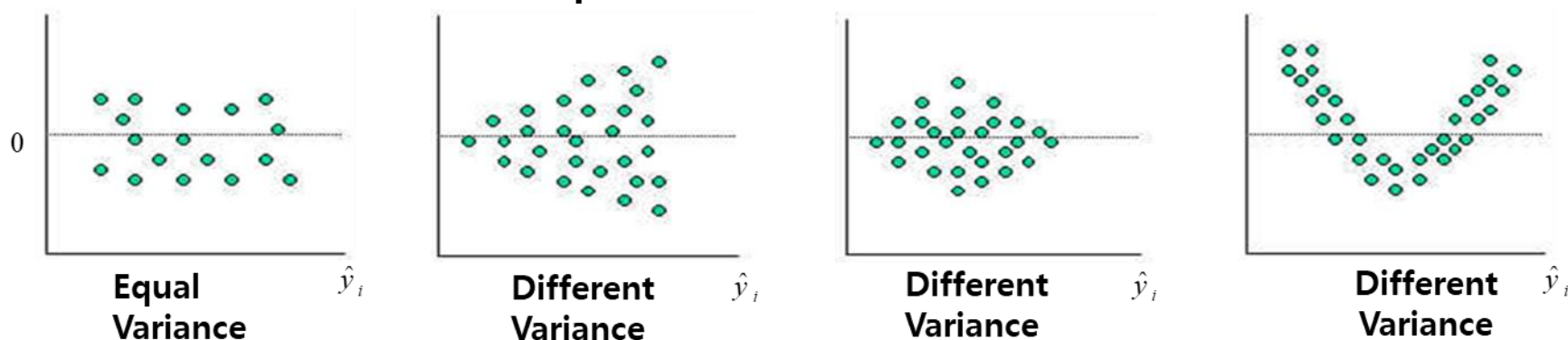
*단순회귀분석에서 y 값의 분포



USING R *잔차(Residual) 분석

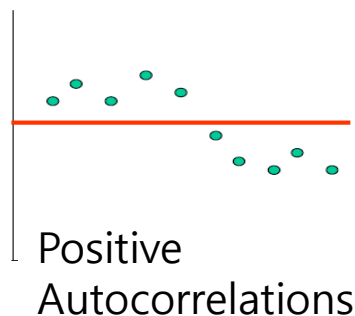
- 종속변수 관측치의 정규성/등분산성을 직접 확인하는 대신, 잔차의 정규성과 등분산성을 확인하는 방법을 주로 사용
- 잔차가 0을 중심으로 특정한 패턴없이 랜덤하게 분포할 때, 등분산성이 만족되었다고 판단
- 잔차의 정규성은 분포의 모양을 확인하거나 Kolmogorov-Smirnov test or Shapiro-Wilk tests로도 확인 가능

Standardized Residual's Scatterplot

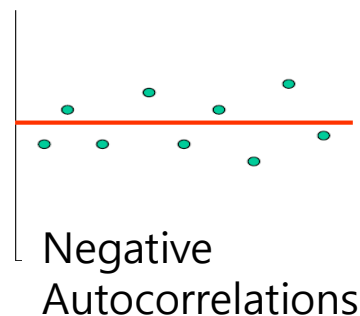


*자기상관(Autocorrelation)

- 자기상관은 연속적인 일련의 관측치들이 서로 상관되어 있을 때 발생한다. 예를 들어, 관측치가 시계열 데이터일 때 데이터는 흔히 높은 자기상관을 갖는다
- 자기상관 확인방법: Durbin-Watson 통계량
 - Durbin-Watson 통계량 = 0: 강한 양의 자기상관
 - Durbin-Watson 통계량 = 2: 자기상관 없음
 - Durbin-Watson 통계량 = 4: 강한 음의 자기상관



Standardized
Residual's
Scatterplot



2. 단순회귀분석의 분석 방법

```
# 회귀선 도표 작성 및 회귀모델의 가정 확인 (정규성, 등분산성, 독립성)
plot(spssdata$self.confidence, spssdata$self.esteem)
abline(regression1_1)
library(sjPlot)
set_theme(axis.title.size = 1.0, axis.textsize = 1.0)
plot_model(regression1_1, type = "diag")
library(lmtest) #to check autocorrelation
dwtest(regression1_1)
```

- ✓ plot 함수에 x값과 y값을 입력
- ✓ abline 함수에 lm함수의 결과값(regression1_1)을 입력하여 회귀선 추가
- ✓ 회귀분석의 가정을 확인하기 위해 plot_model을 활용
 - set_theme 함수로 도표에 출력될 글자 크기를 조정
 - lm 함수로 분석한 회귀분석의 결과가 할당된 객체(regression1_1)를 plot_model 함수에 입력하고 가정 검정 옵션을 지정(type = "diag")
- ✓ 독립성 가정은 Durbin-Watson test로 확인
 - 2를 크게 벗어나면 자기상관이 존재



.허정민

1일 전

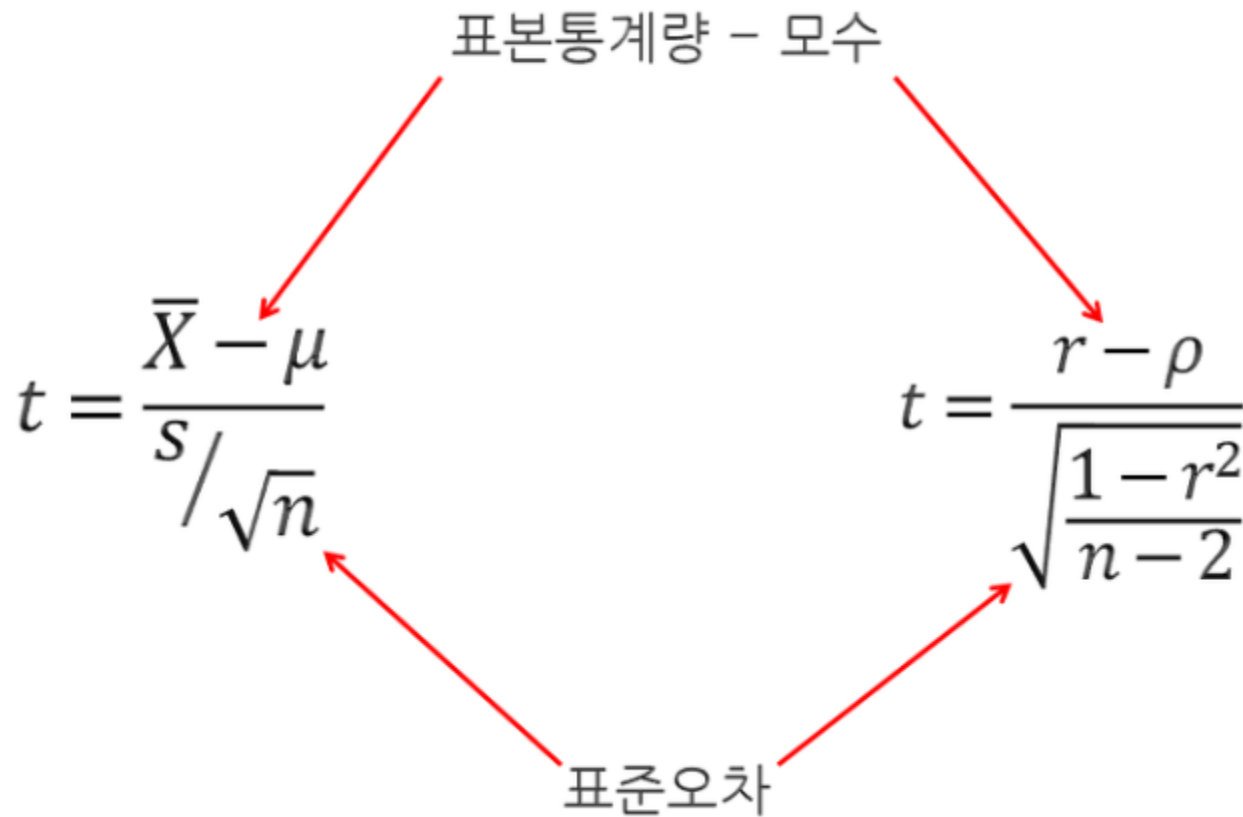
상관계수를 해석할때 궁금한 점이 있습니다!

과제를 하다가 생긴 궁금증인데...

상관 계수를 해석 하는데 있어서 혹시, 그것을 유의미하게 해석할 수 있는 필요한 최소 관측치의 수가 어느정도 될까요?특히 이번처럼 국가에 대한 데이터를 가지고 분석을 할때, 국가가 많아도 200개국인데 본 표본들을 가지고 통계를 진행 하면, 결과가 신뢰로울 수 있는지 궁금합니다.

또한 최소표본이 test마다 다른 이유도 궁금합니다. 제가 알기로는 T,F,chi test 같은 경우 필요한 표본이 다른데 특별한 이유가 있는지도 궁금합니다!

댓글



Information Systems Research

Articles in Advance, pp. 1–12

ISSN 1047-7047 (print) | ISSN 1526-5536 (online)



<http://dx.doi.org/10.1287/isre.2013.0480>

© 2013 INFORMS

Too Big to Fail: Large Samples and the p -Value Problem

Mingfeng Lin

Eller College of Management, University of Arizona, Tucson, Arizona 85721, mingfeng@eller.arizona.edu

Henry C. Lucas, Jr.

Robert Smith School of Business, University of Maryland, College Park, Maryland 20742, hucas@rhsmith.umd.edu

Galit Shmueli

Srini Raju Centre for IT & the Networked Economy, Indian School of Business, Hyderabad 500 032, India, galit.shmueli@isb.edu



.김동원

22시간 전

attach한 package를 detach 할 시, 의존성 있는 package들도 함께 detach 하는 방법이 있나요? 🙏

```
> library(statnet)
필요한 패키지를 로딩중입니다: tergm
필요한 패키지를 로딩중입니다: ergm
필요한 패키지를 로딩중입니다: network
network: Classes for Relational Data
```

이와 같이 원하는 패키지를 attach하면, 패키지에 따라 상이하지만 의존성에 의해 다른 패키지들도 함께 attach되는 경우들이 존재합니다.

수업시간에 배웠던 'statnet'의 경우 해당 패키지를 attach하면 'tergm', 'ergm', 'network', 'networkDynamic', 'ergm.count', 'sna', 'statnet.common', 'tsna' 등의 패키지들이 함께 attach됨을 콘솔창을 통해 확인할 수 있습니다.

이후에 다른 패키지를 실행하려고 하는 경우, 패키지에 빌트인 되어있는 함수들, 예제 데이터셋 등의 이름이 겹치는 문제가 발생하는 경우가 존재하여 이를 detach하고 다시 이용하려고 합니다.

따라서 detach를 이용하여 해당 패키지를 detach 한뒤, search를 통해 확인을 해보면,

```
> detach(package:statnet, unload=TRUE)
> search()
[1] ".GlobalEnv"          "package:tsna"          "package:sna"           "package:statnet.common"
[5] "package:ergm.count"   "package:tergm"         "package:networkDynamic" "package:ergm"
[9] "package:network"      "tools:rstudio"         "package:stats"         "package:graphics"
[13] "package:grDevices"    "package:utils"         "package:datasets"      "package:methods"
[17] "AutoLoads"           "package:base"
```



.한상혁

15시간 전에 게시됨 (10시간 전에 최종 수정됨)

댓글: **attach한 package를 detach 할 시, 의존성 있는 package들도 함께 detach 하는 방법이 있나요?**

해당 내용 관련하여 구글, CRAN을 통해 검색해본 결과 일단 제가 검색한 결과로는 해당 기능을 가진 패키지를 찾지 못하여 해당 기능을 가진 함수를 직접 만들어 봤습니다.

```
package_detacher <- function(input_package){
  ap <- search() # attached packages
  ap <- ap[grepl("package", ap)]
  ap <- gsub("package:", "", ap)
  if(!require(tools)){
    install.packages("tools")
    library(tools)
  }

  padd <- append(package_dependencies(packages=input_package, recursive=T)[[input_package]], input_package) # package and its
  dependencies to detach

  op <- c(setdiff(ap, padd), "stats", "graphics", "grDevices", "utils", "datasets", "methods") # other packages
  opd <- unlist(package_dependencies(packages=op, recursive=T)) # other packages and dependencies
  opd <- c(unique(opd), op)

  is_tools_included <- F

  if("tools" %in% opd){
```




.김다인

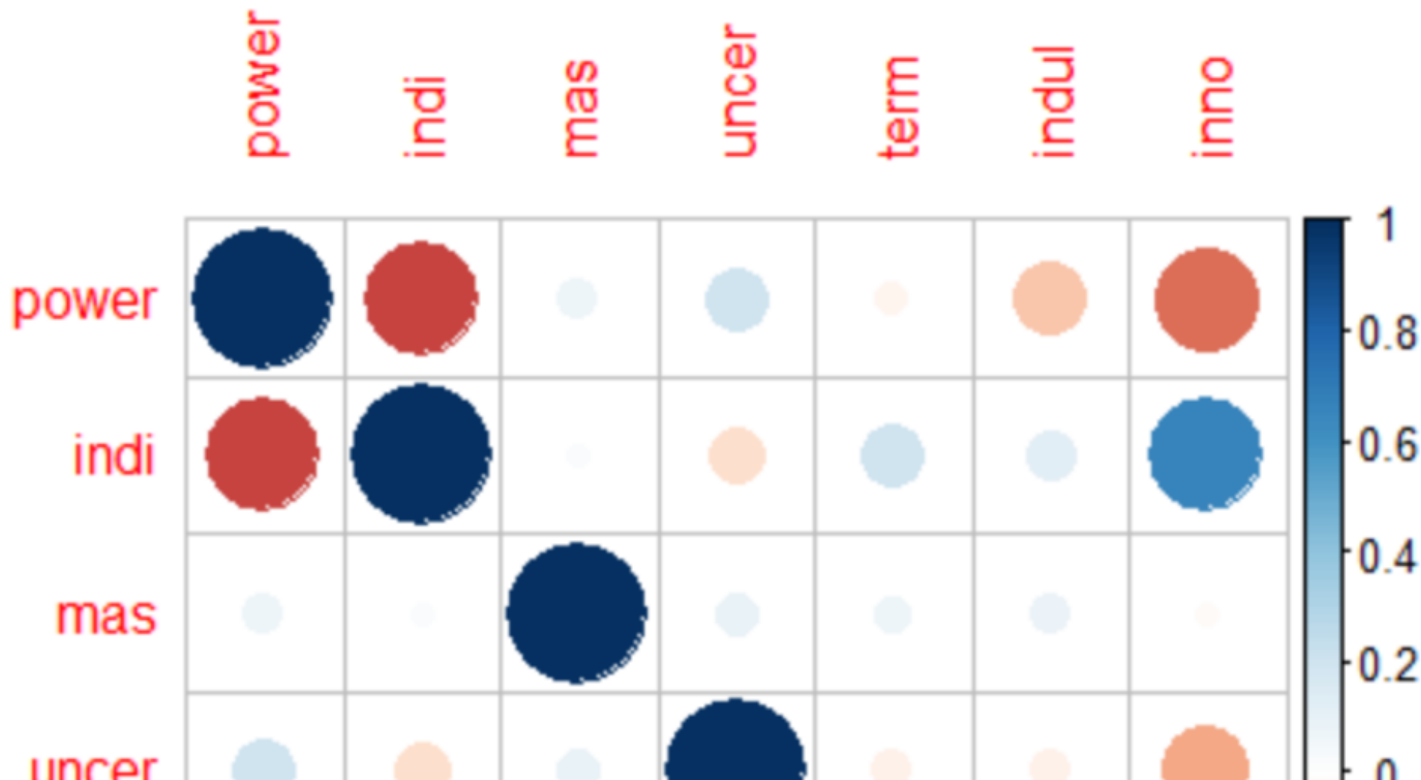
1일 전

corrplot 패키지로 다양하게 시각화하기

밑에 한상혁님이 올려주신 글을 보고 패키지에 대해 찾아보고 corrplot 패키지가 다양한 시각화가 가능하다는 걸 알았습니다. 그래서 추가적으로 글을 올립니다

```
library(corrplot)
```

```
corrplot(cor.var2) #원
```





.김다인

상관계수를 시각화하는 또다른 패키지 **corrgram**

1일 전

```
install.packages("corrgram")
```

```
library(corrgram)
```

한번의 corrgram 사용으로 두 가지 방식의 시각화가 가능합니다.

```
corrgram(데이터, order=TRUE, lower.panel=panel.shade, upper.panel=panel.pie, text.panel=panel.txt,)
```





.김재문

1일 전

회귀분석 결과에서 SST,SSE,SSR, R^2 의 의미

수업시간에 다룬 SST, SSE, SSR, R^2 의 의미에 대하여

실제로 그래프와 평균, 관측 값들을 비교하면서 이해하면 도움이 될 것 같아 시각적으로 정리되어 있는 사이트를 올립니다.

<https://igija.tistory.com/256>

강의 노트와 달리

SSE = Explained Sum of Squares(회귀식으로 설명된 변동)

SSR = Residual Sum of Squares(회귀식으로 설명이 안된 변동)으로

정의가 다르니 참고하시고 보시면 좋을 것 같습니다.

댓글



.한상혁

1일 전

댓글: 회귀분석 결과에서 SST,SSE,SSR, R^2 의 의미

참고로 강의노트 기준으로는 SSE = sum of squared estimate of errors, SSR = sum of squares due to regression 입니다.



.권용재

15시간 전

상관분석시 결측치(missing value)에 대한 해결방법!

cor() 함수는 NA 값이 포함되는 순간 그 결과값은 무조건 NA가 나옵니다.

결측치가 존재하는 경우 2가지 방법을 쓸 수 있습니다.

1. 결측치(NA) 가 존재하는 데이터 row를 삭제하는 방법
> cor(data , use = "complete.obs")
2. 결측치(NA) 가 존재하는 위치에서의 연산만 넘어가는 방법
> cor(data, use= "pairwise.complete.obs")

출처: <https://bioinformaticsandme.tistory.com/59>

댓글



.한상혁

10시간 전

데이터프레임에서 특정 데이터 형식의 열들만 서브세팅하는 간단한 방법

dplyr의 select_if 함수를 사용하시면 간단하게 데이터프레임에서 특정 데이터 형식으로 되어있는 열들만 서브세팅하실 수 있습니다.

```
library(dplyr)
```

```
> df <- data.frame(aa = c("a", "a", "b", "b"), bb = c(1, 2, 3, 4), cc = c("ab", "cd", "ef", "gh"))
```

```
> df
```

```
aa bb cc
```

```
1 a 1 ab
```

```
2 a 2 cd
```

```
3 b 3 ef
```

```
4 b 4 gh
```

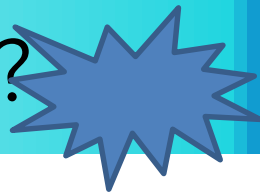
```
> df %>% select_if(is.numeric)
```

```
bb
```

```
1 1
```



*가정이 만족되지 않을 경우, 대안은?



1. 데이터 정제(cleaning)

- Outlier (수정? 제거?)

2. 데이터 변환(transformation)

- Box-cox/로그/지수 변환

3. 일반화선형모델(Generalized linear model)

- (count variable) Poisson regression or Negative binomial regression
- (binary variable) Logistic regression

2. 단순회귀분석의 분석 방법

1. 연구가설 2의 검증

<분석 순서>

- 1) 부모에 대한 애착과 자아존중감도 이미 만들어진 변수가 있기에 그대로 사용한다.
- 2) 부모에 대한 애착이 자아존중감에 미치는 영향에 대한 단순회귀 분석을 수행한다.
- 3) 'sjPlot' 패키지를 이용한 회귀분석 결과표와 도표를 출력한다.

② 연구가설 2의 검증

```
regression1_2 <- lm(self.esteem ~ attachment, data=spssdata)  
summary(regression1_2)
```

- ✓ 연구가설 2는 독립변수가 부모에 대한 애착, 종속변수가 자아존중감
- ✓ 회귀분석을 수행한 결과를 regression1_2라는 객체에 저장하고, summary 함수에 분석 결과를 저장한 객체를 지정하여 회귀분석 결과를 출력

2. 단순회귀분석의 분석 방법

```
> regression1_2 <- lm(self.esteem ~ attachment, data=spssdata)
> summary(regression1_2)
```

```
Call:
lm(formula = self.esteem ~ attachment, data = spssdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-12.9790	-2.4036	-0.0787	2.1208	12.2716

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	14.8026	0.6427	23.033	< 2e-16 ***
attachment	0.2251	0.0309	7.285	1.03e-12 ***

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

Residual standard error: 3.519 on 593 degrees of freedom

Multiple R-squared: 0.08214, Adjusted R-squared: 0.08059

F-statistic: 53.07 on 1 and 593 DF, p-value: 1.032e-12

- ✓ 절편은 14.8026이고, 독립변수의 회귀 계수는 0.2251
- ✓ 부모에 대한 애착이 한 단위 증가할수록 자아존중감은 0.2251만큼 증가

2. 단순회귀분석의 분석 방법

- ✓ 독립변수의 기울기에 대한 t 값은 7.285이고, 유의도는 $1.03e-12$
 - 유의도는 0.05보다 낮은 수준이므로 부모에 대한 애착이 자아존중감에 영향을 미치지 않는다는 영가설은 기각되고, 부모에 대한 애착이 자아존중감에 영향을 미치고 있다는 연구가설이 채택
- ✓ F 값은 53.07이고, F 값에 따른 유의도는 $1.032e-12$ 로 0.05보다 낮은 수준으로 나타나고있어 '회귀모형은 의미가 없다'는 영가설은 기각되고, '회귀모형은 의미가 있다'는 연구가설이 채택
- ✓ 독립변수가 설명하는 종속변수의 분산인 R^2 값은 0.08214로 종속변수의 전체 분산 중에서 8.214%를 독립변수가 설명

2. 단순회귀분석의 분석 방법

```
# ③ 'sjPlot' 패키지를 이용한 회귀분석 결과표와 도표 출력
# 연구가설 1과 연구가설 2의 결과를 하나의 표로 출력
# viewer에 직접 출력하는 방법
library(sjPlot)
tab_model(regression1_1, regression1_2, show.se = T, show.ci = F)
# 결과표를 외부 파일로 저장하는 방법
tab_model(regression1_1, regression1_2, show.se = T, show.ci = F,
pred.labels = c("(Intercept)", "자기신뢰감", "부모에 대한 애착"),
dv.labels = c("자 아 존 중 감 ", "자 아 존 중 감 "), file =
"simple_regression2.html")
file.show("simple_regression2.html")
```

- ✓ 'sjPlot' 패키지에서 plot_model함수는 표 형태로 회귀분석 결과를 출력
 - plot_model 함수는 하나의 회귀모형에 대한 분석 결과를 출력해 줄 수 있을 뿐만 아니라, 여러 개의 회귀모형에 대한 분석 결과를 하나의 표에 출력해 줄 수도 있음
- ✓ plot_model 함수는 lm 함수의 결과를이용하기 때문에 plot_model 함수에는 앞서 연구가설1과 연구가설 2의 회귀분석 결과를 저장한 regression1_1과 regression1_2라는 객체를 입력

2. 단순회귀분석의 분석 방법

<i>Predictors</i>	자아존중감			자아존중감		
	<i>Estimates</i>	<i>std. Error</i>	<i>p</i>	<i>Estimates</i>	<i>std. Error</i>	<i>p</i>
(Intercept)	13.08	0.72	<0.001	14.80	0.64	<0.001
자기신뢰감	0.59	0.07	<0.001			
부모에 대한 애착				0.23	0.03	<0.001
Observations	595			595		
R ² / adjusted R ²	0.117 / 0.116			0.082 / 0.081		

- ✓ 첫 번째 회귀모형에서는 연구가설 1에 대한 검증으로 자아존중감에 대한 자기신뢰감의 영향을 검증한 결과
- ✓ 두 번째 회귀모형은 연구가설 2인 부모에 대한 애착이 자아존중감에 미치는 영향에 대하여 검증한 결과

2. 단순회귀분석의 분석 방법

```
# 연구가설 2에 대한 회귀선 도표 작성 및 회귀분석 가정 확인
plot(spssdata$attachment, spssdata$self.esteem)
abline(regression1_2)
set_theme(axis.title.size = 1.0, axis.textsize = 1.0)
plot_model(regression1_2, type = "diag")
library(lmtest) #to check autocorrelation
dwtest(regression1_2)
```

- ✓ plot 함수에 x값과 y값을 입력
- ✓ abline 함수에 lm함수의 결과값(regression1_2)을 입력하여 회귀선 추가
- ✓ 회귀분석 결과를 도표의 형식으로 출력하기 위해 sjp.lm 함수를 이용
- ✓ set_theme 함수로 도표에 출력될 글자 크기를 조정
- ✓ lm 함수로 분석한 회귀분석의 결과가 할당된 객체(regression1_2)를 plot_model 함수에 입력하고 가정 검정 옵션을 설정(type = "diag")
- ✓ plot_model 함수는 tab_model 함수와 달리 여러 회귀분석에 대한 결과를 함께 출력할 수 없으며, 하나의 회귀분석에 대한 결과만을 출력



3. 다중회귀분석의 분석 방법

연구가설

- 1-1) 성별은 자아존중감에 영향을 미칠 것이다.
- 1-2) 자기신뢰감은 자아존중감에 영향을 미칠 것이다.
- 1-3) 부모에 대한 애착은 자아존중감에 영향을 미칠 것이다.
- 1-4) 부모 감독은 자아존중감에 영향을 미칠 것이다.
- 1-5) 부정적 양육은 자아존중감에 영향을 미칠 것이다.

1. 연구가설 1의 검증

<분석 순서>

- 1) 분석에 필요한 독립변수를 만든다.
- 2) 다중회귀분석을 수행한다.
- 3) R은 SPSS와 달리 표준화된 계수를 자동적으로 계산해주지 않기 때문에 별도로 표준화된 계수를 구하는 작업이 필요하다.
- 4) 독립변수들 간의 다중공선성 여부를 진단한다.
- 5) 'sjPlot' 패키지를 이용해서 결과표 및 도표를 출력한다.

3. 다중회귀분석의 분석 방법

```
# ① 변수 만들기
# 부모 감독 변수 만들기
attach(spssdata)
spssdata$monitor <- q33a07w1+q33a08w1+q33a09w1+q33a10w1
# 성별 변수 재부호화
spssdata$sexw1.re[sexw1 == 1] <- 0 #shorter version?
spssdata$sexw1.re[sexw1 == 2] <- 1
detach(spssdata)
```

- ✓ 연구가설 1에서 종속변수는 자아존중감(self-esteem)이고, 독립변수는 성별(sexw1), 자기신뢰감(self.confidence), 부모에 대한 애착(attachment), 부모 감독(monitor), 그리고 부정적 양육(negative.parenting)으로 총 5개
- ✓ 이들 변수들 중에서 부모 감독을 제외하고 나머지 변수들은 이미 만들었던 변수들이므로 부모 감독을 만들면 다중회귀분석을 시행
- ✓ 성별 변수(sexw1)는 남자 청소년인 경우에는 '1'로, 여자 청소년인 경우에는 '2'로 측정되었다. 성별 변수의 변수값을 남자 청소년인 경우에는 '0'으로, 여자 청소년인 경우에는 '1'로 재부호화한 변수(sexw1.re)를 만들어 **남자 청소년을 기준으로 여자 청소년인 경우의 영향을 살펴해보도록 함**

3. 다중회귀분석의 분석 방법

② 다중회귀분석

```
regression2 <- lm(self.esteem ~ sexw1.re+self.confidence+
  attachment+monitor+negative.parenting, data=spssdata)
summary(regression2)
```

- ✓ lm 함수에는 종속변수와 독립변수 순서로 입력
- ✓ 종속변수와 독립변수 사이에 '~' 기호를 입력하고, 독립변수들 사이에는 '+' 기호를 입력하고, 종속변수와 독립변수가 있는 데이터를 지정
- ✓ lm 함수를 이용한 회귀분석의 결과는 regression2라는 객체에 저장하고, summary 함수를 사용하여 회귀분석의 결과를 확인

```
> regression2 <- lm(self.esteem ~ sexw1.re+self.confidence+
+ attachment+ monitor+negative.parenting, data=spssdata)
> summary(regression2)
Call:
lm(formula = self.esteem ~ sexw1.re + self.confidence + attachment
  +monitor + negative.parenting, data = spssdata)
```

3. 다중회귀분석의 분석 방법

Residuals:

Min	1Q	Median	3Q	Max
-12.210	-1.993	0.003	1.875	13.196

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	11.32468	0.98859	11.455	< 2e-16	***
sexw1.re	-0.28566	0.28015	-1.020	0.3083	
self.confidence	0.50306	0.06729	7.476	2.79e-13	***
attachment	0.16252	0.03472	4.681	3.54e-06	***
monitor	0.02294	0.04924	0.466	0.6415	
negative.parenting	-0.11532	0.04532	-2.545	0.0112	*

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

Residual standard error: 3.335 on 588 degrees of freedom
(1 observation deleted due to missingness)

Multiple R-squared: 0.1823, Adjusted R-squared: 0.1754
F-statistic: 26.23 on 5 and 588 DF, p-value: < 2.2e-16

- ✓ 성별의 회귀 계수는 -0.28566, 자기신뢰감은 0.50306, 부모에 대한 애착은 0.16252, 부모 감독은 0.02294, 그리고 부정적 양육은 -0.11532



3. 다중회귀분석의 분석 방법

- ✓ 성별과 부모 감독은 유의도가 0.05보다 **높게** 나타나고 있어 성별과 부모 감독은 종속변수에 영향을 미치지 않는다는 **영가설**을 채택
- ✓ 자기신뢰감, 부모에 대한 애착, 그리고 부정적 양육은 유의도가 모두 0.05보다 **낮은** 수준으로 나타남
 - 이들 독립변수는 종속변수에 대해 통계적으로 유의한 영향을 미치고 있음
- ✓ F 값이 26.23로 유의도는 $2.2e-16$ 으로 나타남
 - F 값에 따른 유의도가 0.05보다 낮은 수준이므로 회귀모형은 의미가 없다는 영가설을 기각하고, **회귀모형은 의미가 있다는** 연구가설을 채택
- ✓ R^2 값은 0.1823로 종속변수의 분산 중에서 독립변수에 의해 설명되는 분산은 18.23%

*sexw1을 쓸 때와 sexw1.re를 쓸 때의 차이

```
#Compare sexw1 with sexw1.re
regression2_1 <- lm(self.esteem ~ sexw1+self.confidence+
                    attachment+monitor+negative.parenting, data =
spssdata)
summary(regression2_1)
regression2$coefficients
regression2_1$coefficients #any difference?
```

3. 다중회귀분석의 분석 방법

```
# ③ 표준화 계수값 구하기
install.packages("QuantPsyc")
library(QuantPsyc)
lm.beta(regression2)
round(lm.beta(regression2), 3) # 표준화 계수값을 소수점 3자리로 지정
```

- ✓ 표준화 계수는 'QuantPsyc' 패키지의 lm.beta 함수를 통해 구함
 - lm.beta 함수에는 다중회귀분석의 결과를 저장한 객체를 입력
 - round 함수를 이용하여 표준화 계수의 소수점을 지정

```
> lm.beta(regression2)
sexw1.re self.confidence attachment monitor negative.parenting
-0.04202828 0.30069582 0.19993262 0.02322585 -0.10692453
> round(lm.beta(regression2), 3)
sexw1.re self.confidence attachment monitor negative.parenting
-0.042 0.301 0.200 0.023 -0.107
```

- ✓ 자기신뢰감의 표준화 계수가 0.301로 가장 높게 나타나고 있어, 자아 존중감에 가장 큰 영향을 미치는 독립변수는 자기신뢰감

3. 다중회귀분석의 분석 방법

```
# ④ 다중공선성 진단을 위한 vif 값 구하기
install.packages("car")
library(car)
vif(regression2)
vif(regression2) < 10
```

- ✓ 분산팽창계수는 'car' 패키지의 vif 함수를 이용해서 구함
 - vif 함수는 다중회귀분석의 결과를 저장한 객체를 입력하여 구함
 - vif 값이 클수록 다중공선성으로 인해 회귀모형이 문제가 있다는 것을 의미함

```
> vif(regression2)
sexw1.re self.confidence attachment monitor negative.parenting
1.044100      1.107805      1.395480 1.405715      1.062209
```

- ✓ 모든 독립변수의 vif 값이 2 미만으로 나타나고 있어 다중공선성으로 인한 문제는 없음

$$VIF_i = \frac{1}{1 - R_i^2}$$

- 만약 vif 값이 10 이상인 변수가 존재한다면 다중공선성의 문제를 해결하기 위해서 해당 변수를 제외하거나, 해당 변수와 유사한 다른 변수와 더하여 하나의 변수로 만드는 방법을 사용

3. 다중회귀분석의 분석 방법

```
# ⑤ 'sjPlot' 패키지를 이용한 결과표 및 도표 출력
# 다중회귀분석 결과표 작성
tab_model(regression2, show.se = T, show.ci = F, show.stat= T)
# 한글을 사용
tab_model(regression2, show.se = T, show.ci = F, show.stat= T,
pred.labels = c("(Intercept)", "성별", "자기신뢰감", "부모에 대한 애착",
"부모 감독", "부정적 양육"), dv.labels = c("자아존중감"), file =
"multiple_regression.html")
file.show("multiple_regression.html")
```

- ✓ 'sjPlot' 패키지를 이용하여 다중회귀분석의 결과를 표나 도표의 형태로 출력
 - tab_model 함수는 다중회귀분석의 결과를 표의 형식으로 출력할 수 있는 함수
 - tab_model 함수에 lm 함수에서 다중회귀분석 모형의 결과를 저장한 객체를 지정
 - 결과표에 출력될 통계량을 몇몇 인자를 이용하여 지정

3. 다중회귀분석의 분석 방법

자아존중감				
<i>Predictors</i>	<i>Estimates</i>	<i>std. Error</i>	<i>Statistic</i>	<i>p</i>
(Intercept)	11.32	0.99	11.46	<0.001
성별	-0.29	0.28	-1.02	0.308
자기신뢰감	0.50	0.07	7.48	<0.001
부모에 대한 애착	0.16	0.03	4.68	<0.001
부모 감독	0.02	0.05	0.47	0.641
부정적 양육	-0.12	0.05	-2.54	0.011
Observations	594			
R ² / adjusted R ²	0.182 / 0.175			

- ✓ 다섯 개의 독립변수에 대한 앞서 지정한 통계량이 출력됨
 - 회귀 계수(B), 표준오차(std.Error), t값(statistic), 그리고 유의도(p)로 출력됨
 - 추가로 분석에 사용한 사례수(Observations), R² 값과 수정된 R² 값(R²/adj. R²)이 출력됨

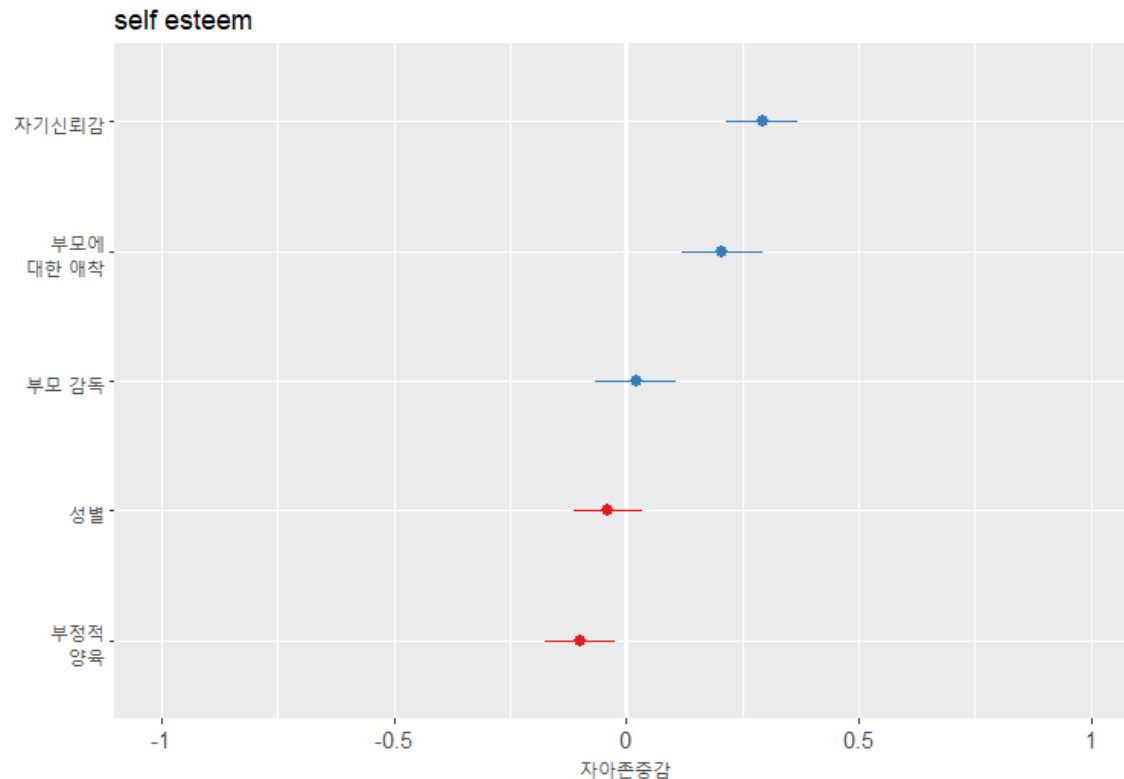
3. 다중회귀분석의 분석 방법

```
# 'sjPlot' 패키지를 이용한 다중회귀분석 도표 작성
set_theme(axis.title.size = 1.0, axis.textsize = 1.0)
plot_model(regression2, type = "est", wrap.labels=5)
plot_model(regression2, type = "est", axis.labels=c("성별", "자기신
뢰감", "부정적 양육", "부모 감독", "부모에 대한 애착"), axis.title="자아
존중감", wrap.labels=5)
# 'sjPlot' 패키지를 이용한 다중회귀분석 도표에 표준화 계수를 이용하여 작성
plot_model(regression2, type = "std", sort.est = T, wrap.labels=5)
plot_model(regression2, type = "std", sort.est = T,
axis.labels=c("부정적 양육", "성별", "부모 감독", "부모에 대한 애착", "자
기신뢰감"), axis.title="자아존중감", wrap.labels=5)
# 'sjPlot' 패키지를 이용한 다중공선성 진단 및 회귀분석 가정 검정
plot_model(regression2, type = "diag")
library(lmtest)
dwtest(regression2)
```

- ✓ plot_model 함수에는 다중회귀분석의 결과를 저장한 객체 (regression2)를 지정하고, y축에 출력될 변수 설명을 label에서 가져오는 대신 lm 함수에 입력한 **독립변수의 순서대로** 직접 입력 (axis.labels=c("성별", "자기신뢰감", "부모에 대한 애착", "부모 감독", "부정적 양육"), auto.label = F)

3. 다중회귀분석의 분석 방법

- ✓ 만일 회귀 계수 대신 표준화 계수를 표시하고자 한다면 `type = "std"` 옵션을 이용하여 도표를 출력

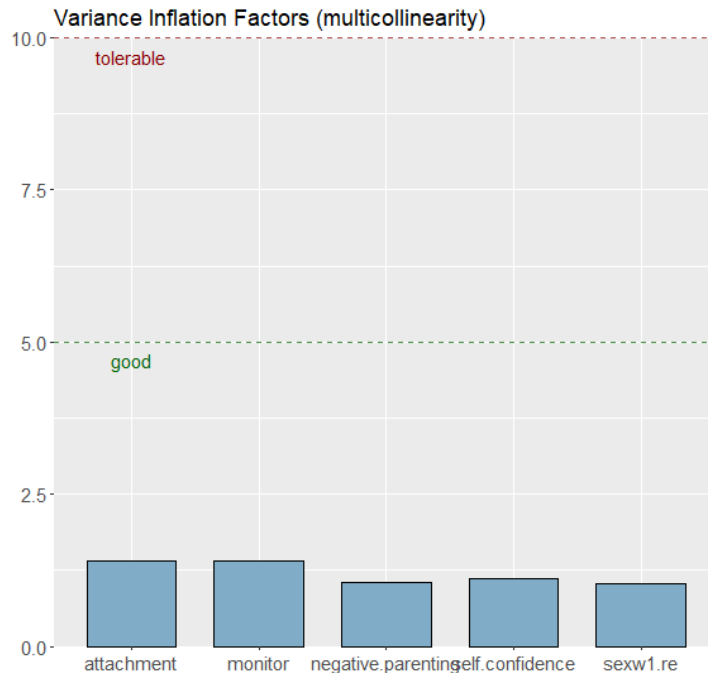


- ✓ 회귀 계수가 음수인 경우에는 빨간색(본서에서는 파란색) 점과 숫자로 표시되고, 양수인 경우에는 파란색(본서에서는 검정색)으로 표시됨
 - 선의 길이는 신뢰구간을 의미함



3. 다중회귀분석의 분석 방법

- ✓ 다중회귀모형에 대한 다중공선성 진단의 결과 및 회귀분석 기본 가정 만족여부를 도표로 출력하기 위해서는 type 인자를 "diag"로 지정



- ✓ 다중공선성 진단 결과를 살펴보면, 독립변수들의 vif 값이 모두 2.5 이하인 것을 확인할 수 있음
 - 다중공선성으로 인한 문제는 없는 것으로 판단



.김동원

3일 전

RE: attach한 package를 detach 할 시, 의존성 있는 package들도 함께 detach 하는 방법이 있나요?

Facebook R community에 해당 질문을 공유하여 받은 답변을 추가로 기재합니다.

남성현

패키지를 detach하셔야 하는 구체적인 상황이 어떤 것인지 모르겠으나, clean R session이 필요하신 거라면, `here::here()` 사용하시는 코딩습관을 들이시면 좋을 것 같은데요. 관련 글을 링크합니다. 관련없는 것이라면 죄송합니다.

<https://malco.io/2018/11/05/why-should-i-use-the-here-package-when-i-m-already-using-projects/>

이승섭

한패키지의 의존성 패키지 리스트만 알면 되는것같은데요.. 어차피 dete~ 하는 기능은 똑같으니깐요.. 패키지 설치시 의존성을 잘 모르면 사용하기 힘들어서 이번에 한번 찾아보셔도.. 깔면 이제 libdev 같은 os 파일들 까지.. 쿨럭

패키지 내리는 경우가 종종 발생하는 경우는 사용함수가 패키지끼리 충돌되어 예러가 나거나 취사선택하게 되는데 이게 사용자 의도와 다를 수 있습니다. 패키지명을 함수앞에 ::넣어도 한방법이긴 한데 그래도 되지않는 경우가 있더라구용 이경우에도 찾아보면 있긴하던데..

ex rpostgresql 과 sqldf를 등록하여 sqldf를 날리면 dataframe을 인풋으로 잡아도 자꾸 pg디비에 날라가는 현상이..

패키지를 detech한다는 거 자체가 사용기능을 좀더 명확하고 메모리 부하가 적어 좋은 습관인것 같습니다 물론 그전에 쓸 패키지만 올리는게 좋지만요

댓글

인용

수정

삭제

이메일

3. 다중회귀분석의 분석 방법



2. 모형 선택 방법(예측 중심)

<분석 순서>

- 1) 데이터에서 결측값 사례를 제외한다.
- 2) 전진 선택 방법에 사용하기 위해서 모든 독립변수를 투입한 회귀 분석 결과와 기본 모형의 결과를 객체에 할당한다
- 3) step 함수를 이용하여 전진 선택 방법의 모형 선택을 분석한다.

1) 전진 선택 방법

① 결측값 사례 제외

```
spssdata.no.na <- na.omit(spssdata[c("self.esteem", "sexw1.re",  
  "self.confidence", "attachment", "monitor",  
  "negative.parenting")])
```

- ✓ 분석에 앞서 독립변수와 종속변수에 결측값이 없는 데이터를 만들기
 - 전진 선택 방법이나 후진 선택 방법에서 변수가 추가되거나 제거 되는데, 투입되거나 제거되는 변수에 따라 결측값 사례수가 다르게 되면 분석 결과의 일관성이 떨어지기 때문

3. 다중회귀분석의 분석 방법

② 전체 모형과 기본 모형

```
regression3 <- lm(self.esteem ~ sexw1.re+self.confidence+
  attachment+monitor+negative.parenting, data=spssdata.no.na)
null <- lm(self.esteem ~ 1, data=spssdata.no.na)
```

③ 전진 선택 모형

```
step1 <- step(null, scope=list(lower=null, upper=regression3),
  direction="forward")
summary(step1)
```

- ✓ 전진 선택 방법에 사용하기 위해서 모든 독립변수를 투입한 회귀분석 결과와 기본 모형의 결과를 각각 객체에 할당
 - 전진 선택 방법을 통한 모형 선택을 분석하기 위해 모든 독립변수를 투입한 **전체 모형을 regression3**이라는 객체에 할당 (모두 입력하는 대신 "**self.esteem ~ .**"으로 표현하는 것도 가능)
 - 이 분석에서는 **독립변수가 전혀 없는 모형**을 기본 모형으로 가정하여 **null**이라는 객체에 저장
- ✓ step 함수를 이용한 모형 선택에서 전진 선택 방법을 적용
 - scope 인자를 통해 기본 모형(lower=null)과 연구자가 지정한 회귀 모형(upper=regression3)을 입력

3. 다중회귀분석의 분석 방법

```
> spssdata.no.na <- na.omit(spssdata[c("self.esteem", "sexw1.re",
"self.confidence", "attachment", "monitor", "negative.parenting")])
> regression3 <- lm(self.esteem~
sexw1.re+attachment+self.confidence+monitor+negative.parenting,
data = spssdata.no.na)
> step1 <- step(null, scope = list(lower = null, upper =
regression3), direction = "forward")
Start: AIC=1546.53
self.esteem ~ 1
```

	Df	Sum of Sq	RSS	AIC
+ self.confidence	1	952.13	7047.1	1473.2
+ attachment	1	662.56	7336.6	1497.2
+ monitor	1	372.76	7626.4	1520.2
+ negative.parenting	1	226.97	7772.2	1531.4
<none>			7999.2	1546.5
+ sexw1.re	1	3.33	7995.9	1548.3

- ✓ 첫 번째 단계에서는 기본 모형이 적용되었고, 기본 모형에서 **각 독립 변수가 추가되었을 경우의** 자유도(Df), 제곱합(Sum of Sq), 잔차 제곱합(RSS: Residual Sum of Square), 그리고 AIC 값이 출력됨

3. 다중회귀분석의 분석 방법

- ✓ AIC 값을 기준으로 살펴보면, 독립변수를 전혀 가정하지 않은 기본 모형에 자기신뢰감 변수(self.confidence)가 추가되었을 경우의 AIC 값은 1473.2
- ✓ 자기신뢰감 변수가 추가되었을 경우에는 **기본 모형의 AIC 값보다 낮고, 다른 변수가 추가되었을 때보다 AIC 값이 더 낮기 때문에** 첫 번째 단계에서 선택된 독립변수는 자기신뢰감

3. 다중회귀분석의 분석 방법

Step: AIC=1473.25
self.esteem ~ self.confidence

	Df	Sum of Sq	RSS	AIC
+ attachment	1	418.84	6628.2	1438.8
+ negative.parenting	1	174.84	6872.2	1460.3
+ monitor	1	122.47	6924.6	1464.8
<none>			7047.1	1473.2
+ sexw1.re	1	0.07	7047.0	1475.2

- ✓ 두 번째 단계에서의 모형은 이전 단계에서 선택된 독립변수인 자기 신뢰감이 추가된 모형임
- ✓ 독립변수가 추가되었을 경우에 기본 모형의 AIC 값보다 낮은 AIC 값을 갖는 독립변수는 부모에 대한 애착(attachment), 부정적 양육(negative.parenting), 그리고 부모 감독(monitor)
- ✓ 이들 변수 중에서 **AIC 값이 가장 낮은 독립변수는 부모에 대한 애착** 이므로 두 번째 단계에서 선택된 독립변수는 부모에 대한 애착

3. 다중회귀분석의 분석 방법

Step: AIC=1438.85

self.esteem ~ self.confidence + attachment

	Df	Sum of Sq	RSS	AIC
+ negative.parenting	1	74.529	6553.7	1434.1
<none>			6628.2	1438.8
+ sexw1.re	1	11.947	6616.3	1439.8
+ monitor	1	2.484	6625.7	1440.6

- ✓ 세 번째 단계에서의 모형은 독립변수로 자기신뢰감과 부모에 대한 애착을 가정한 모형이 됨
- ✓ 이전 단계에서와 같은 방법으로 살펴본 결과, **부정적 양육이 세 번째 단계에서 선택됨**

3. 다중회귀분석의 분석 방법

Step: AIC=1434.14

self.esteem ~ self.confidence + attachment + negative.parenting

	Df	Sum of Sq	RSS	AIC
<none>			6553.7	1434.1
+ sexw1.re	1	10.6813	6543.0	1435.2
+ monitor	1	1.5304	6552.2	1436.0

- ✓ 네 번째 단계에서는 독립변수로 자기신뢰감, 부모에 대한 애착, 그리고 부정적 양육이 가정된 기본 모형의 AIC 값보다 기본 모형에 추가되었을 경우에 **AIC 값이 낮아지는 독립변수가 없으므로 전진 선택 방법은 중단됨**

*Deviance, AIC, BIC

$$1. \textit{Deviance} = 2(\ln(\widehat{L}_S) - \ln(\widehat{L})) = -2 \ln(\widehat{L})$$

- \widehat{L}_S : 포화 모델의 Likelihood function의 최대값은 1, 그러므로 $\ln(\widehat{L}_S) = 0$
- \widehat{L} : 모델의 Likelihood function의 최대값

$$2. \textit{AIC} = 2k - 2 \ln(\widehat{L})$$

- Akaike information criterion
- k : 모델에서 추정해야 하는 모수의 개수

$$3. \textit{BIC} = \ln(n)k - 2 \ln(\widehat{L})$$

- Bayesian information criterion
- n : 샘플 데이터 개수(크기)

3. 다중회귀분석의 분석 방법

```
> summary(step1)
```

```
Call:
```

```
lm(formula = self.esteem ~ self.confidence + attachment +  
negative.parenting, data = spssdata.no.na)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-12.273	-2.025	0.015	1.975	13.191

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.31481	0.96920	11.674	< 2e-16 ***
self.confidence	0.51675	0.06508	7.940	1.02e-14 ***
attachment	0.16416	0.03066	5.355	1.23e-07 ***
negative.parenting	-0.11717	0.04524	-2.590	0.00983 **

```
---
```

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

```
Residual standard error: 3.333 on 590 degrees of freedom
```

```
Multiple R-squared:  0.1807,    Adjusted R-squared:  0.1765
```

```
F-statistic: 43.38 on 3 and 590 DF,  p-value: < 2.2e-16
```



3. 다중회귀분석의 분석 방법

- ✓ 전진 선택 방법을 통한 최종 모형은 독립변수로 자기신뢰감, 부모에 대한 애착, 그리고 부정적 양육인 포함된 회귀 모형
- ✓ **최종 모형의 결과는 summary 함수에 step 함수를 이용한 결과를 저장한 객체(step1)를 입력하면 확인할 수 있음**
 - 독립변수인 자기신뢰감, 부모에 대한 애착, 그리고 부정적 양육은 종속변수인 자아존중감에 모두 통계적으로 유의한 영향을 미치고 있음
- ✓ 최종 모형의 F 값은 43.38이고, 유의도는 $2.2e-16$ 으로 0.05보다 낮은 수준이므로 최종 모형은 의미가 있는 것으로 나타남
- ✓ R^2 값은 0.1807로 종속변수인 자아존중감의 전체 분산 중에서 3개의 독립변수가 18.07%를 설명하고 있음

3. 다중회귀분석의 분석 방법

2) 후진 선택 방법

<분석 순서>

- 1) 데이터에서 결측값 사례를 제외한다.
- 2) 후진 제거 방법에 사용하기 위해서 모든 독립변수를 투입한 회귀 분석 결과를 객체에 할당한다
- 3) step 함수를 이용하여 후진 제거 방법의 모형 선택을 분석한다.

③ 후진 제거 모형

```
step2 <- step(regression3, direction="backward")
summary(step2)
```

- ✓ 후진 제거 방법은 step 함수에서 연구자가 가정한 회귀 모형을 입력한 후에 direction 인자를 backward로 지정
- ✓ 연구자가 가정한 회귀 모형에서 모든 독립변수를 투입한 회귀분석 결과를 할당한 객체(regression3)를 지정
- ✓ 후진 제거 방법을 통해 얻은 결과를 출력하기 위해 step 함수의 내용을 step2라는 객체에 할당하고, summary 함수를 통해 결과를 출력

3. 다중회귀분석의 분석 방법

```
> step2 <- step(regression3, direction = "backward")
Start: AIC=1436.95
self.esteem ~ sexw1.re + attachment + self.confidence + monitor +
negative.parenting
```

	Df	Sum of Sq	RSS	AIC
- monitor	1	2.41	6543.0	1435.2
- sexw1.re	1	11.57	6552.2	1436.0
<none>			6540.6	1437.0
- negative.parenting	1	72.03	6612.6	1441.5
- attachment	1	243.76	6784.4	1456.7
- self.confidence	1	621.71	7162.3	1488.9

- ✓ 첫 번째 단계에서는 사용자가 지정한 회귀 모형에서 모든 독립변수를 가정한 회귀 모형이 기본 모형이 됨
- ✓ 기본 모형에서 각 독립변수를 제외했을 경우에 예상되는 AIC 값을 살펴보면, 부모 감독(monitor)의 AIC 값은 1435.2로 기본 모형일 경우보다 낮게 나타났고, 다른 독립변수를 제외했을 경우보다 가장 낮게 나타남
- ✓ 첫 번째 단계에서 제거될 독립변수는 부모감독이 됨

3. 다중회귀분석의 분석 방법

Step: AIC=1435.17

```
self.esteem ~ sexw1.re + attachment + self.confidence +
negative.parenting
```

	Df	Sum of Sq	RSS	AIC
- sexw1.re	1	10.68	6553.7	1434.1
<none>			6543.0	1435.2
- negative.parenting	1	73.26	6616.3	1439.8
- attachment	1	329.20	6872.2	1462.3
- self.confidence	1	676.27	7219.3	1491.6

- ✓ 두 번째 단계에서는 첫 번째 단계에서 제외된 부모 감독을 제외한 회귀 모형을 기본 모형으로 가정
- ✓ 두 번째 단계의 기본 모형에서 각 독립변수를 제외했을 경우의 AIC 값을 비교해 보면, 성별(sexw1.re)이 제외될 경우에 AIC 값은 1434.1로 기본 모형의 AIC 값(1435.17)보다 더 낮아짐
- ✓ 성별을 제외하는 것이 더 간명성이 있는 회귀 모형인 것으로 나타남

3. 다중회귀분석의 분석 방법

Step: AIC=1434.14

self.esteem ~ attachment + self.confidence + negative.parenting

	Df	Sum of Sq	RSS	AIC
<none>			6553.7	1434.1
- negative.parenting	1	74.53	6628.2	1438.8
- attachment	1	318.52	6872.2	1460.3
- self.confidence	1	700.37	7254.1	1492.5

- ✓ 세 번째 단계에서는 첫 번째와 두 번째 단계에서 제외된 부모 감독과 성별을 제외한 회귀모형이 기본 모형이 되고, AIC 값은 1434.14
- ✓ 세 번째 단계에서는 독립변수를 제외할 경우에 기본 모형보다 AIC 값이 더 낮게 나타나지 않으므로 후진 제거 방법은 중단
- 후진 제거 방법의 최종 모형
 - ✓ 후진 제거 방법을 통해 선택한 모형은 부정적 양육, 부모에 대한 애착, 그리고 자기신뢰감이 독립변수로 가정된 모형
 - ✓ 후진 제거 방법을 통해 얻은 회귀 분석 결과는 step 함수의 내용을 저장한 객체인 step2를 summary 함수에 입력하면 확인할 수 있음



3. 다중회귀분석의 분석 방법

3) 선택 제거 방법

<분석 순서>

- 1) 데이터에서 결측값 사례를 제외한다.
- 2) 선택 제거 방법에 사용하기 위해서 모든 독립변수를 투입한 회귀 분석 결과를 객체에 할당한다
- 3) step 함수를 이용하여 선택 제거 방법의 모형 선택을 분석한다.

③ 선택 제거 방법

```
step3 <- step(regression3, direction="both")  
summary(step3)
```

- ✓ step 함수에서 사용자가 가정한 회귀 모형을 입력하고, direction 인자에 both로 지정
- ✓ step 함수의 내용을 step3이라는 객체에 저장하여 선택 제거 방법을 통해 얻은 회귀 모형의 결과를 확인할 수 있도록 함

3. 다중회귀분석의 분석 방법

```
> step3 <- step(regression3, direction = "both")
Start: AIC=1436.95
self.esteem ~ sexw1.re + attachment + self.confidence + monitor +
negative.parenting
```

	Df	Sum of Sq	RSS	AIC
- monitor	1	2.41	6543.0	1435.2
- sexw1.re	1	11.57	6552.2	1436.0
<none>			6540.6	1437.0
- negative.parenting	1	72.03	6612.6	1441.5
- attachment	1	243.76	6784.4	1456.7
- self.confidence	1	621.71	7162.3	1488.9

- ✓ 선택 제거 방법의 첫 번째 단계에서는 후진 제거 방법과 같이 사용자가 회귀 모형에서 지정한 **모든 독립변수를 가정한 모형이 기본 모형**
- ✓ 기본 모형에서 각 독립변수를 제거했을 경우에 예상되는 AIC 값을 비교해 보면, 부모 감독(monitor)을 제거했을 경우의 AIC 값이 1435.2로 기본 모형의 AIC 값보다 낮고, 다른 어떤 독립변수를 제거했을 경우보다 가장 낮게 나타남
- ✓ 첫 번째 단계에서 제거될 독립변수는 부모 감독

3. 다중회귀분석의 분석 방법

Step: AIC=1435.17

```
self.esteem ~ sexw1.re + attachment + self.confidence +
negative.parenting
```

	Df	Sum of Sq	RSS	AIC
- sexw1.re	1	10.68	6553.7	1434.1
<none>			6543.0	1435.2
+ monitor	1	2.41	6540.6	1437.0
- negative.parenting	1	73.26	6616.3	1439.8
- attachment	1	329.20	6872.2	1462.3
- self.confidence	1	676.27	7219.3	1491.6

- ✓ 두 번째 단계에서는 첫 번째 단계에서 제거된 부모 감독을 제외한 회귀 모형이 기본 모형
- ✓ 기본 모형에서 각 독립변수를 제거했을 경우의 AIC 값이 출력되고, 첫 번째 단계에서 제거되었던 부모 감독이 다시 추가되었을 경우의 AIC 값도 함께 출력됨
- ✓ 선택 제거 방법은 후진 제거 방법을 통해 각 단계 마다 독립변수를 제거하지만, 동시에 이전 단계에서 제거된 독립변수를 다시 추가했을 경우의 AIC 값을 계산하여 기본 모형과 비교

3. 다중회귀분석의 분석 방법

- ✓ 이전 단계에서 제거된 독립변수가 다시 추가된 경우의 통계량은 변수명 앞에 '+' 표시가 되고, 각 단계의 기본 모형에서 제거되었을 경우의 통계량은 변수명 앞에 '-' 표시가 되어 구분됨
- ✓ 첫 번째 단계에서 제거되었던 부모 감독을 다시 추가한 모형은 첫 번째 단계의 기본 모형과 같은 모형이 되므로 AIC 값은 첫 번째 단계 기본 모형의 AIC 값과 같은 1437.0이 되고, 두 번째 기본 모형의 AIC 값보다 크기 때문에 부모 감독은 두 번째 단계에서 추가되지 않는 것이 더 좋은 모형이 됨
- ✓ 기본 모형에서 성별(sexw1.re)을 제거했을 경우의 AIC 값은 1434.1로 기본 모형의 AIC 값보다 더 낮은 것으로 나타나고 있음
- ✓ 두 번째 단계의 결과는 첫 번째 단계에서 제거되었던 부모 감독은 다시 추가되지 않고, 두 번째 기본 모형에서 성별을 제거하는 것이 더 좋은 모형이 되는 것으로 나타남

3. 다중회귀분석의 분석 방법

Step: AIC=1434.14

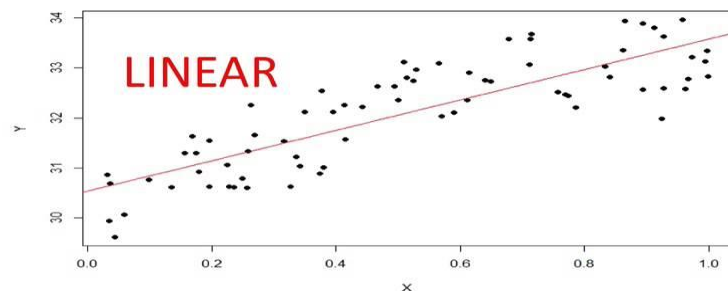
self.esteem ~ attachment + self.confidence + negative.parenting

	Df	Sum of Sq	RSS	AIC
<none>			6553.7	1434.1
+ sexw1.re	1	10.68	6543.0	1435.2
+ monitor	1	1.53	6552.2	1436.0
- negative.parenting	1	74.53	6628.2	1438.8
- attachment	1	318.52	6872.2	1460.3
- self.confidence	1	700.37	7254.1	1492.5

- ✓ 세 번째 단계에서는 이전 두 단계에서 부모 감독과 성별을 제외한 회귀 모형이 기본 모형
- ✓ 기본 모형에서 이전 두 단계에서 제거된 성별이나 부모감독이 다시 추가되었을 경우의 AIC 값은 각각 1435.2와 1436.0으로 세 번째 단계 기본 모형의 AIC 값보다 더 높게 나타남
- ✓ 세 번째 단계에서 부모 감독이나 성별은 다시 추가되지 않는 것이 더 좋은 모형
- ✓ 세 번째 기본 모형에서 더 이상 제거될 독립변수가 없는 것을 의미하고, 모형 선택 과정은 종료

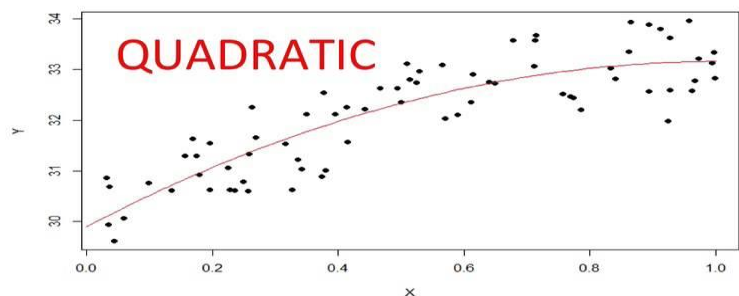
3. 다중회귀분석의 분석 방법

- 선택 제거 방법의 최종 모형
 - ✓ 선택 제거 방법을 통해 선택한 모형은 부정적 양육, 부모에 대한 애착, 그리고 자기신뢰감이 독립변수로 포함하는 모형
 - ✓ 선택 제거 방법을 통해 얻은 회귀 분석 결과는 step 함수의 내용을 저장한 객체인 step3를 summary 함수를 사용해서 확인할 수 있음



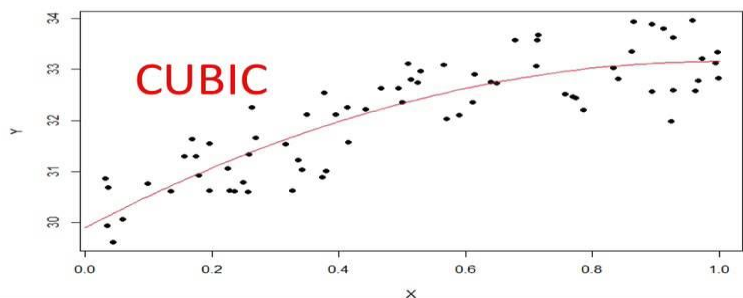
Multiple R-squared: 0.7044

$$Y = 30.53 + 3.05 * X$$



Multiple R-squared: 0.7559

$$Y = 29.90 + 6.48 * X - 3.22 * X^2$$



Multiple R-squared: 0.7623

$$Y = 30.17 + 3.61 * X + 3.71 * X^2 - 4.48 * X^3$$

```
model <- lm(y ~ x + I(x^2) + I(x^3))
```

```
혹은 model <- lm(y ~ poly(x,3))
```



```
## Polynomial regression
plot(spssdata$self.confidence, spssdata$self.esteem,
main="Polynomial Regression")
poly1 <- lm(self.esteem ~ self.confidence, data = spssdata)
summary(poly1)
abline(poly1, lwd=3, col="red")

# first, the WRONG WAY...
poly2 <- lm(self.esteem ~ self.confidence + self.confidence^2,
data = spssdata)
summary(poly2)

# now, the RIGHT WAY...
poly2 <- lm(self.esteem ~ self.confidence + I(self.confidence^2),
data = spssdata)
summary(poly2)
spssdata$self.confidence2 <- spssdata$self.confidence^2
poly2_1 <- lm(self.esteem ~ self.confidence + self.confidence2,
data = spssdata)
summary(poly2_1)
poly2_2 <- lm(self.esteem ~ poly(x = self.confidence, degree = 2,
raw = T), data = spssdata)
summary(poly2_2)

# compare these two models by partial F test
anova(poly1, poly2)
```






3. 다중회귀분석의 분석 방법

3. 더미변수를 이용한 회귀분석의 분석 방법

연구가설

- 1-1) 성별은 자아존중감에 영향을 미칠 것이다.
- 1-2) 부모에 대한 애착은 자아존중감에 영향을 미칠 것이다.
- 1-3) 부모 감독은 자아존중감에 영향을 미칠 것이다.
- 1-4) 직업 결정 상태는 자아존중감에 영향을 미칠 것이다.

<분석 순서>

- 1) 집단 변수(직업 결정 상태)를 더미변수로 만든다.
- 2) 더미변수를 이용한 다중회귀분석을 수행한다.
- 3) 'sjPlot' 패키지를 이용한 다중회귀분석 결과표를 작성한다.



3. 다중회귀분석의 분석 방법

1) 더미변수를 만들어 연구가설을 검증하는 방법

```
# ① 더미변수 만들기
spssdata$job.dummy1 <- ifelse(spssdata$q2w1==2, 1, 0)
spssdata$job.dummy2 <- ifelse(spssdata$q2w1==3, 1, 0)
#library(sjlabelled)
spssdata$job.dummy1 <- set_label(spssdata$job.dummy1,
                                "직업결정_대강의 생각")
spssdata$job.dummy2 <- set_label(spssdata$job.dummy2,
                                "직업결정_정해지지 않음")

# 더미변수 확인
table(spssdata$q2w1)
table(spssdata$job.dummy1)
table(spssdata$job.dummy2)
```

- ✓ '직업 결정 상태(q2w1)' 변수의 값이 2인 경우에는 1로 재부호화하고, 변수의 값이 2 이외의 값일 경우에는 0으로 재부호화하여 spssdata라는 데이터의 job.dummy1이라는 변수에 저장
- ✓ '직업 결정 상태' 변수의 값이 3인 경우에는 1로 재부호화하고, 변수의 값이 3 이외의 값일 경우에는 0으로 재부호화하여 spssdata라는 데이터의 job.dummy2라는 변수에 저장



3. 다중회귀분석의 분석 방법

- ✓ 더미변수에 'sjlabelled' 패키지의 set_label 함수로 변수 설명을 입력하고, table 함수로 제대로 더미변수가 만들어졌는지 확인

```
> spssdata$job.dummy1 <- ifelse(spssdata$q2w1==2, 1, 0)
> spssdata$job.dummy2 <- ifelse(spssdata$q2w1==3, 1, 0)
> table(spssdata$q2w1)
```

```
  1    2    3
129 353 113
```

```
> table(spssdata$job.dummy1)
```

```
  0    1
242 353
```

```
> table(spssdata$job.dummy2)
```

```
  0    1
482 113
```

3. 다중회귀분석의 분석 방법

```
# ② 더미변수를 이용한 다중회귀분석
regression4 <- lm(self.esteem ~ sexw1.re+attachment+monitor+
  job.dummy1+job.dummy2, data=spssdata)
summary(regression4)

# 표준화 계수 출력
library(QuantPsyc)
lm.beta(regression4)
```

- ✓ lm 함수에는 연구가설에 따른 종속변수와 독립변수를 차례로 입력하고, 해당 변수들이 있는 데이터를 지정
- ✓ 다중회귀분석의 결과를 regression4라는 객체에 저장한다. 분석 결과를 살펴보기 위해 summary 함수에 분석 결과가 저장된 객체 (regression4)를 입력

3. 다중회귀분석의 분석 방법

```
> regression4 <- lm(self.esteem ~ sexw1.re+attachment+monitor+
+ job.dummy1+job.dummy2, data=spssdata)
> summary(regression4)
```

Call:

```
lm(formula = self.esteem ~ sexw1.re + attachment + monitor +
job.dummy1 + job.dummy2, data = spssdata)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-13.6872	-2.4654	0.7161	2.6271	12.5168

Coefficients:

	Estimate	Std. Error	t value	pr(> t)
(Intercept)	15.38092	0.79284	19.400	< 2e-16 ***
sexw1.re	-0.62211	0.29167	-2.133	0.03334 *
attachment	0.18659	0.03546	5.262	2e-07 ***
monitor	0.10230	0.05010	2.042	0.04160 *
job.dummy1	-0.92307	0.36063	-2.560	0.01073 *
job.dummy2	-1.44868	0.45403	-3.191	0.00149 **

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

Residual standard error: 3.469 on 587 degrees of freedom
(2 observations deleted due to missingness)

Multiple R-squared: 0.1109, Adjusted R-squared: 0.1033

F-statistic: 14.64 on 5 and 587 DF, p-value: 1.544e-13

3. 다중회귀분석의 분석 방법

```
> lm.beta(regression4)
      sexw1.re attachment      monitor  job.dummy1  job.dummy2
-0.08497708  0.23794181  0.09240798 -0.12378117 -0.15437845
```

- ✓ 다중회귀분석의 결과를 살펴보면, 성별, 부모에 대한 애착, 부모 감독, 그리고 두 개의 더미변수 모두 종속변수인 자아존중감에 통계적으로 유의한 영향을 미치고 있는 것으로 나타남
- ✓ **더미변수의 준거집단은 '(1) 구체적으로 확정해 놓은 직업이 있다'**
- ✓ 첫 번째 더미변수(job.dummy1)에서 변수값이 1로 재부호화한 집단은 '(2) 확정적이지는 않지만 대강 생각해 놓은 직업이 있다'
 - 더미변수에서 변수값이 1인 집단('(2) 확정적이지는 않지만 대강 생각해 놓은 직업이 있다')이 준거집단('(1) 구체적으로 확정해 놓은 직업이 있다')에 비해 **자아존중감이 낮은 것으로 해석**
- ✓ 두 번째 더미변수의 영향에 대해서도 변수값이 1인 집단('(3) 아직 정해놓은 장래의 직업이 없다')이 준거집단에 비해 **자아존중감이 낮은 것으로 해석**

3. 다중회귀분석의 분석 방법

```
# ③ 'sjPlot' 패키지를 이용한 다중회귀분석 결과표 작성
tab_model(regression4, show.se = T, show.ci = F, show.stat= T,
auto.label = F)
tab_model(regression4, show.se = T, show.ci = F, show.stat= T,
pred.labels = c("(Intercept)", "성별", "부모에 대한 애착", "부모 감독",
"직업결정_대강의 생각", "직업결정_정해지지 않음"), dv.labels = c("자아존중
감"), file = "multiple_regression2.html")
file.show("multiple_regression2.html")
```

✓ 'sjPlot' 패키지의 tab_model 함수를 이용하여 회귀분석의 결과를 출력

<i>Predictors</i>	자아존중감			
	<i>Estimates</i>	<i>std. Error</i>	<i>Statistic</i>	<i>p</i>
(Intercept)	15.38	0.79	19.40	<0.001
성별	-0.62	0.29	-2.13	0.033
부모에 대한 애착	0.19	0.04	5.26	<0.001
부모 감독	0.10	0.05	2.04	0.042
직업결정_대강의 생각	-0.92	0.36	-2.56	0.011
직업결정_정해지지 않음	-1.45	0.45	-3.19	0.001
Observations	593			
R ² / adjusted R ²	0.111 / 0.103			



3. 다중회귀분석의 분석 방법

2) factor 변수로 변환하여 연구가설을 검증하는 방법

```
# ① 더미변수로 만들 변수를 factor 변수로 변환하기  
spssdata$q2w1a <- factor(spssdata$q2w1)
```

- ✓ factor 함수는 성별이나 지역과 같은 범주형 변수로 데이터에 저장하는데 사용되는 함수
 - lm 함수는 독립변수를 factor로 인식하게 되면 **해당 변수의 변수 값에서 가장 낮은 값의 집단을 준거집단으로 삼아** 일시적으로 더미변수를 만들어 분석
- ✓ 더미변수로 만들 변수를 factor로 인식할 수 있게 만들 수 있는 방법으로는 첫 번째로 해당 변수를 factor 변수로 만드는 것
 - 이 방법은 factor 함수를 이용하여 변환할 수 있음
 - factor 함수에 대상 변수를 입력하고, **factor로 변환된 새로운 변수를 객체(q2w1a)에 저장**

3. 다중회귀분석의 분석 방법

```
# ② factor 변수가 포함된 다중회귀분석
# factor 변수로 다중회귀분석 시행
regression5 <- lm(self.esteem ~ sexw1.re+attachment+monitor+q2w1a,
  data=spssdata)
summary(regression5)
```

- ✓ lm 함수에는 종속변수와 독립변수를 차례대로 입력하고, 변수가 있는 데이터를 지정
- ✓ 독립변수 중 더미변수로 분석할 변수는 factor로 변환한 객체(q2w1a)를 대신 입력하여 분석

```
> spssdata$q2w1 <- to_factor(spssdata$q2w1)
> regression5 <- lm(self.esteem ~ sexw1.re+attachment+monitor+
+ q2w1a, data=spssdata)
> summary(regression5)
Call:
lm(formula = self.esteem ~ sexw1.re + attachment + monitor + q2w1a,
data = spssdata)

Residuals:
    Min       1Q   Median       3Q      Max
-13.6872  -2.4654  -0.7161   2.6271  12.5168
```

3. 다중회귀분석의 분석 방법

Coefficients:

	Estimate	Std. Error	t value	pr(> t)	
(Intercept)	15.38092	0.79284	19.400	< 2e-16	***
sexw1.re	-0.62211	0.29167	-2.133	0.03334	*
attachment	0.18659	0.03546	5.262	2e-07	***
monitor	0.10230	0.05010	2.042	0.04160	*
q2w1a2	-0.92307	0.36063	-2.560	0.01073	*
q2w1a3	-1.44868	0.45403	-3.191	0.00149	**

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

Residual standard error: 3.469 on 587 degrees of freedom
 (2 observations deleted due to missingness)
 Multiple R-squared: 0.1109, Adjusted R-squared: 0.1033
 F-statistic: 14.64 on 5 and 587 DF, p-value: 1.544e-13

- ✓ 더미변수는 q2w1a2와 q2w1a3로 출력
- ✓ 각각 더미변수의 원변수인 q2w1a에서 변수값 2번을 1로 재부호화한 더미변수(q2w1a2)와 변수값 3번을 1로 재부호화한 더미변수(q2w1a3)

3. 다중회귀분석의 분석 방법

```
# as.factor 함수로 더미변수가 포함된 다중회귀분석 결과 산출
regression5_1 <- lm(self.esteem ~ sexw1.re+attachment+monitor+
as.factor(q2w1), data=spssdata)
summary(regression5_1)
```

- ✓ 더미변수로 만들어야 하는 변수를 factor로 lm 함수에서 인식시킬 수 있는 또 다른 방법으로는 lm 함수에 더미변수로 만들 변수에 대해 **as.factor** 함수로 지정하는 방법
- ✓ lm 함수에는 종속변수와 독립변수를 차례대로 입력하고, 더미변수로 만들 변수를 as.factor에 지정
- ✓ lm 함수의 결과를 regression5_1이라는 객체에 저장하고, summary 함수에 이 객체를 지정하여 분석 결과를 확인
- ✓ 분석 결과에는 더미변수로 변환된 변수이름이 as.factor(q2w1)2와 as.factor(q2w1)3으로 출력됨
- ✓ 이 결과에서도 factor로 인식된 q2w1 변수값 2번을 1로 재부호화한 더미변수와 변수값 3번을 1로 재부호화한 더미변수

3. 다중회귀분석의 분석 방법

```
> regression5_1 <- lm(self.esteem ~ sexw1.re+attachment+
+ monitor+as.factor(q2w1), data=spssdata)
> summary(regression5_1)
```

Call:

```
lm(formula = self.esteem ~ sexw1.re + attachment + monitor +
    as.factor(q2w1), data = spssdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.6872	-2.4654	-0.7161	2.6271	12.5168

Coefficients:

	Estimate	Std. Error	t value	pr(> t)	
(Intercept)	15.38092	0.79284	19.400	< 2e-16	***
sexw1.re	-0.62211	0.29167	-2.133	0.03334	*
attachment	0.18659	0.03546	5.262	2e-07	***
monitor	0.10230	0.05010	2.042	0.04160	*
as.factor(q2w1)2	-0.92307	0.36063	-2.560	0.01073	*
as.factor(q2w1)3	-1.44868	0.45403	-3.191	0.00149	**

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

Residual standard error: 3.469 on 587 degrees of freedom
(2 observations deleted due to missingness)

Multiple R-squared: 0.1109, Adjusted R-squared: 0.1033

F-statistic: 14.64 on 5 and 587 DF, p-value: 1.544e-13

3. 다중회귀분석의 분석 방법

```
# ③ 'sjPlot' 패키지를 이용한 다중회귀분석 결과표 작성
tab_model(regression5, show.se = T, show.ci = F, show.stat= T,
auto.label = F)
tab_model(regression5, show.se = T, show.ci = F, show.stat= T,
pred.labels = c("(Intercept)", "성별", "부모에 대한 애착", "부모 감독",
"직업결정_대강의 생각", "직업결정_정해지지 않음"), dv.labels = c("자아존중
감"), file = "multiple_regression2_1.html")
file.show("multiple_regression2_1.html")
```

- ✓ tab_model 함수에 factor로 인식된 변수로 회귀분석한 결과 (regression5)를 지정하고, 출력할 통계량과 변수에 대한 설명 입력

자아존중감				
<i>Predictors</i>	<i>Estimates</i>		<i>std. Error</i>	<i>Statistic</i>
				<i>p</i>
(Intercept)	15.38	0.79	19.40	<0.001
성별	-0.62	0.29	-2.13	0.033
부모에 대한 애착	0.19	0.04	5.26	<0.001
부모 감독	0.10	0.05	2.04	0.042
직업결정_대강의 생각	-0.92	0.36	-2.56	0.011
직업결정_정해지지 않음	-1.45	0.45	-3.19	0.001
Observations	593			
R ² / adjusted R ²	0.111 / 0.103			

3. 다중회귀분석의 분석 방법



4 독립변수들간의 상호작용 효과를 포함한 회귀분석 방법

연구가설

- 1-1) 성별은 자기신뢰감에 영향을 미칠 것이다.
- 1-2) 부정적 양육은 자기신뢰감에 영향을 미칠 것이다.
- 1-3) 부모에 대한 애착은 자기신뢰감에 영향을 미칠 것이다.
- 1-4) 부모 감독은 자기신뢰감에 영향을 미칠 것이다.
- 1-5) 부모에 대한 애착 정도에 따라
부모 감독이 자기신뢰감에 미치는 영향이 달라질 것이다.

<분석 순서>

- 1) 다중공선성 문제의 해결을 위해서 상호작용 효과를 분석할 새로운 변수를 만든다.
- 2) 독립변수들 간의 상호작용 효과를 포함하는 다중회귀분석을 수행한다.
- 3) 다중회귀분석 모형의 다중공선성을 진단한다.
- 4) 'sjPlot' 패키지를 이용하여 결과표와 도표를 출력한다.

3. 다중회귀분석의 분석 방법

```
# ① 변수 만들기
# 각 변수의 평균을 0으로 만들기
mean.attachment <- mean(spssdata$attachment, na.rm=TRUE)
spssdata$centered.attachment <- spssdata$attachment -
  mean.attachment
mean.monitor <- mean(spssdata$monitor, na.rm=TRUE)
spssdata$centered.monitor <- spssdata$monitor - mean.monitor
```

✓ 독립변수를 평균 중심화(Mean Centering)

- 독립변수와 새로운 상호작용 변수 간에 높은 상관관계로 인해서 나타날 수 있는 다중공선성 문제를 해결하기 위한 방법
- 상호작용 효과를 살펴볼 독립변수의 평균을 0으로 만드는 방법으로 연구자가 직접 변수의 평균을 구하고, 해당 변수에서 평균을 빼는 방법

3. 다중회귀분석의 분석 방법

```
# scale 함수를 이용한 평균 중심화 방법
# scale 함수를 이용하여 각 변수의 평균을 0으로 만들기
spssdata$centered.attachment2 <- scale(spssdata$attachment,
                                       center=TRUE, scale=FALSE)[,1]
spssdata$centered.monitor2 <- scale(spssdata$monitor, center=TRUE,
                                    scale=FALSE)[,1]
# 두 방법 비교
head(cbind(spssdata$centered.attachment,
            spssdata$centered.attachment2))
head(cbind(spssdata$centered.monitor, spssdata$centered.monitor2))

# scale 함수를 이용하여 표준화된 값으로 변환(변수의 평균은 0이고, 분산은 1로 변환)
spssdata$centered.attachment3 <- scale(spssdata$attachment,
                                       center=TRUE, scale=TRUE)[,1]
spssdata$centered.monitor3 <- scale(spssdata$monitor, center=TRUE,
                                    scale=TRUE)[,1]
library(psych)
describe(spssdata$centered.attachment3)
describe(spssdata$centered.monitor3)
```

- ✓ 평균 중심화를 할 수 있는 또 다른 방법은 scale 함수를 이용하는 것
- ✓ scale 함수를 이용하면 상호작용 효과를 살펴볼 독립변수의 평균을 0으로 만들거나, 표준화 시킬 수 있음

3. 다중회귀분석의 분석 방법

```
# ② 변수들 간의 상호작용을 가정한 다중회귀분석
regression6 <- lm(self.confidence ~ sexw1.re+negative.parenting+
                  centered.attachment*centered.monitor, data=spssdata)
summary(regression6)
regression6_1 <- lm(self.confidence ~
sexw1.re+negative.parenting+attachment+monitor, data=spssdata)
summary(regression6_1) #check direct effect
regression6_2 <- lm(self.confidence ~
sexw1.re+negative.parenting+attachment*monitor, data=spssdata)
summary(regression6_2) #interaction term without centering
```

- ✓ 회귀분석의 결과를 출력하기 위해 lm 함수에는 종속변수와 독립변수를 차례대로 입력
- ✓ 상호작용 효과를 살펴보기 위한 변수들 사이에는 '*' 표시를 함
- ✓ 지정된 변수가 있는 데이터를 지정하고, 분석결과를 객체(regression6)에 저장
- ✓ summary함수에 회귀분석 결과를 저장한 객체를 지정하면 결과를 출력할 수 있음

3. 다중회귀분석의 분석 방법

```
> regression6 <- lm(self.confidence ~ sexw1.re+
+ negative.parenting+centered.attachment*centered.monitor,
+ data=spssdata)
> summary(regression6)
```

Call:

```
lm(formula = self.confidence ~ sexw1.re + negative.parenting +
    centered.attachment * centered.monitor, data = spssdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.8748	-1.3320	0.0682	1.3437	5.0757

Coefficients:

	Estimate	Std. Error	t value	pr(> t)
(Intercept)	10.856515	0.221342	49.049	< 2e-16 ***
sexw1.re	-0.467651	0.166274	-2.813	0.00508 **
negative.parenting	-0.004043	0.027188	-0.149	0.88184
centered.attachment	0.029329	0.020786	1.411	0.15878
centered.monitor	0.170767	0.028702	5.950	4.63e-09 ***
centered.attachment:				
centered.monitor	0.011990	0.004312	2.781	0.00560 **

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



3. 다중회귀분석의 분석 방법

Residual standard error: 1.99 on 586 degrees of freedom
(3 observations deleted due to missingness)
Multiple R-squared: 0.109, Adjusted R-squared: 0.1014
F-statistic: 14.33 on 5 and 586 DF, p-value: 2.977e-13

- ✓ 종속변수인, 자기신뢰감에 통계적으로 유의한 영향을 미치는 독립변수로는 성별(sexw1.re), 부모 감독(centered.monitor), 그리고 부모에 대한 애착과 부모 감독의 상호작용(centered.attachmentcentered.monitor)으로 나타남
- ✓ 부모에 대한 애착과 부모 감독의 **상호작용의 회귀 계수는 양수이므로 자기신뢰감에 정적(positive) 영향을 미치고** 있는 것으로 나타남
 - 두 변수의 직접효과(direct effect)는 상호작용항이 없는 상태에서 확인(regression6_1)
- ✓ 구체적으로 부모에 대한 애착과 부모 감독의 상호작용이 자기신뢰감에 미치는 영향의 형태에 대해서는 단순히 회귀 계수를 통해 해석하기보다 **상호작용 그래프를 통해** 자세히 살펴보는 것이 필요함

3. 다중회귀분석의 분석 방법

③ 다중공선성 진단

```
library(car)
vif(regression6)
vif(regression6_2)
```

- ✓ 상호작용 효과를 검증하기 위한 다중회귀분석에서 독립변수들 간의 다중공선성을 진단하기 위해 'car' 패키지의 vif 함수를 이용

```
> vif(regression6)
               sexw1.re
               1.032667
      negative.parenting
               1.062868
    centered.attachment
               1.402413
      centered.monitor
               1.330824
centered.attachment:centered.monitor
               1.011612
```

- ✓ 모든 독립변수와 상호작용 변수에서 vif 값이 모두 2 이내
 - 다중공선성의 문제는 없는 것으로 판단



- ✓ 평균중심화를 하지 않으면 상호작용과 관련된 변수들 간에 높은 vif 값이 발생

```
> vif(regression6_2)
      sexw1.re negative.parenting attachment
      1.032667      1.062868      12.519072
      monitor attachment:monitor
      13.423290      35.634541
```

3. 다중회귀분석의 분석 방법

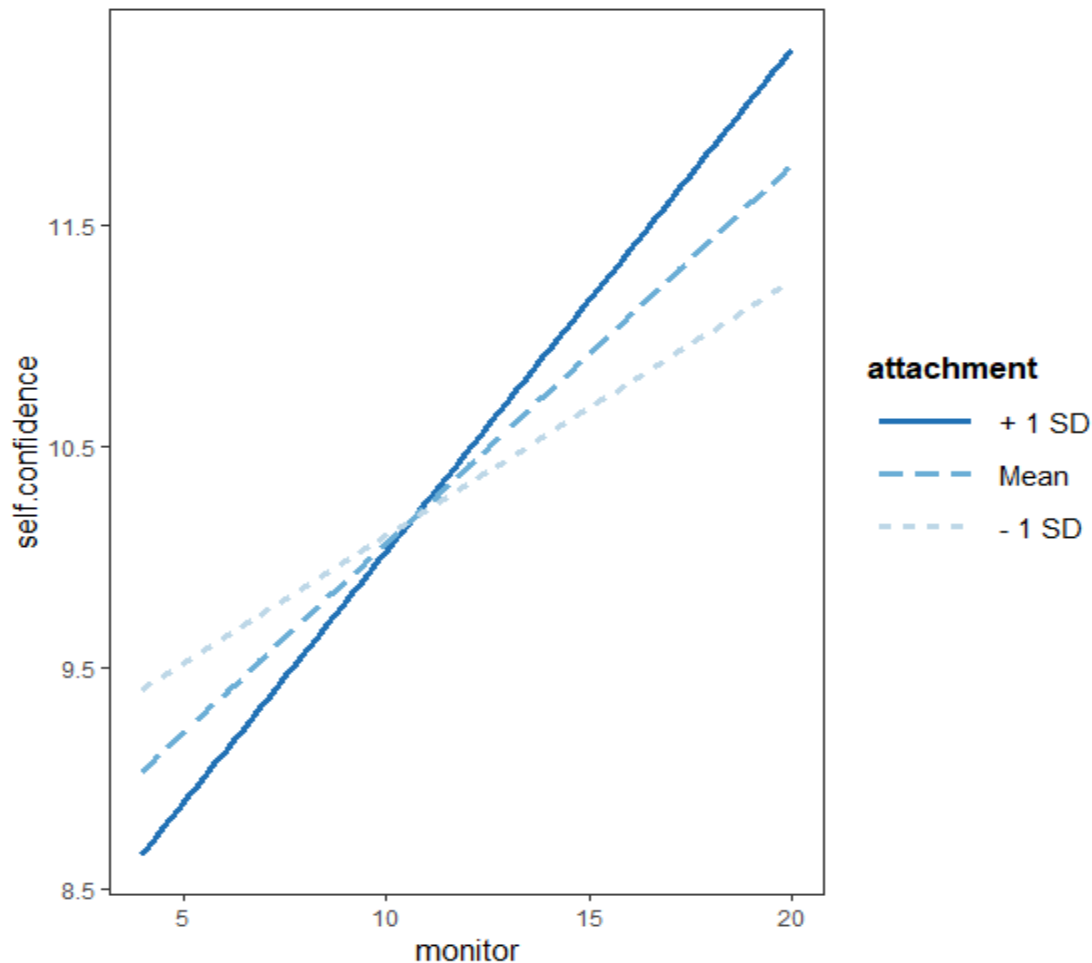
```
# ④ 'sjPlot' 패키지를 이용한 상호작용 도표 작성
# 'sjPlot' 패키지를 이용한 다중회귀분석 결과표 작성
tab_model(regression6, show.se = T, show.ci = F, show.stat= T,
auto.label = F)
tab_model(regression6, show.se = T, show.ci = F, show.stat= T,
pred.labels = c("(Intercept)", "성별", "부정적 양육", "부모에 대한 애착",
"부모 감독", "애착과 감독의 상호작용"), dv.labels = c("자아존중감"),
file = "multiple_regression3.html")
file.show("multiple_regression3.html")
```

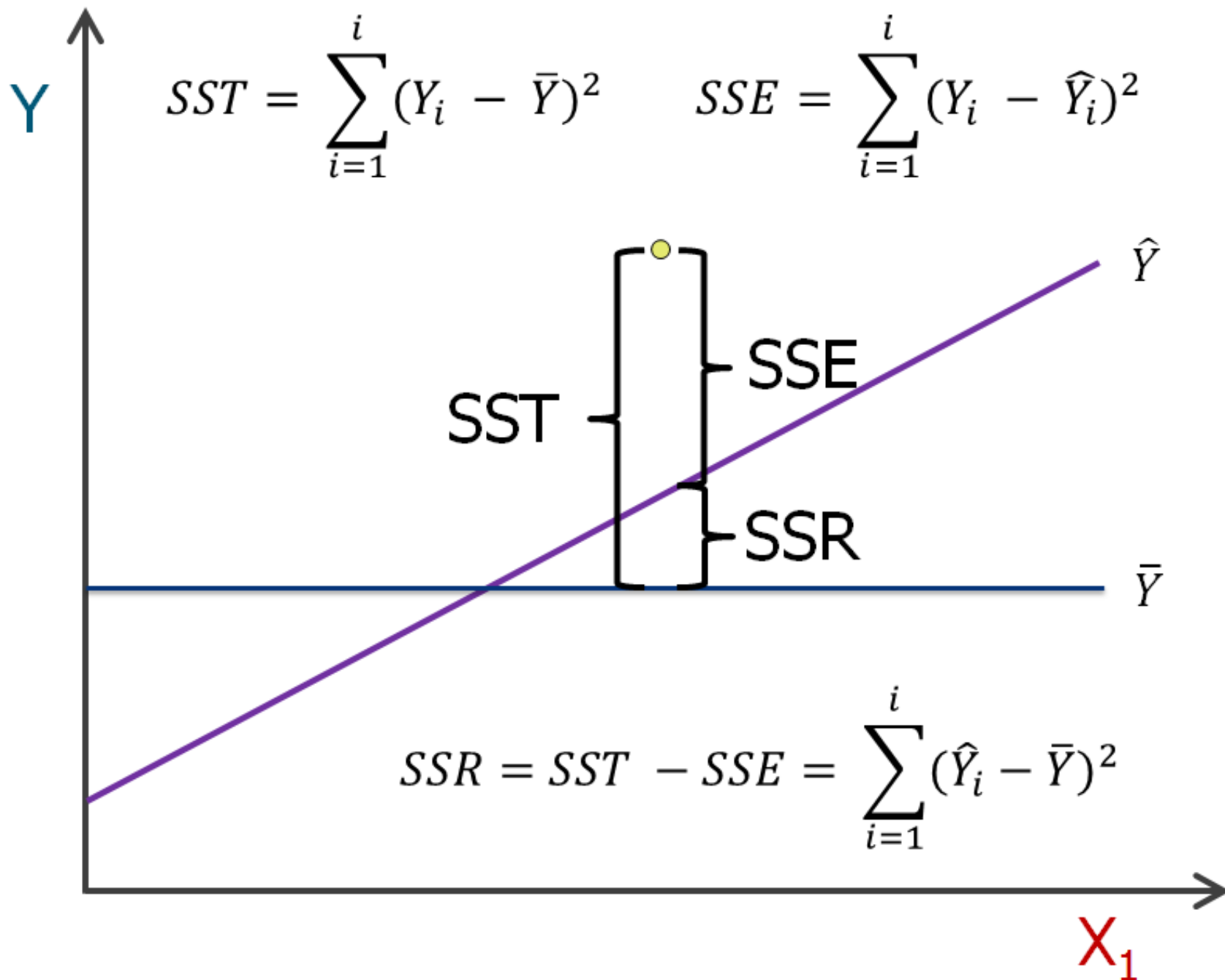
자아존중감				
<i>Predictors</i>	<i>Estimates</i>		<i>std. Error</i>	<i>Statistic</i> <i>p</i>
(Intercept)	10.86	0.22	49.05	<0.001
성별	-0.47	0.17	-2.81	0.005
부정적 양육	-0.00	0.03	-0.15	0.882
부모에 대한 애착	0.03	0.02	1.41	0.159
부모 감독	0.17	0.03	5.95	<0.001
애착과 감독의 상호작용	0.01	0.00	2.78	0.006
Observations	592			
R ² / adjusted R ²	0.109 / 0.101			



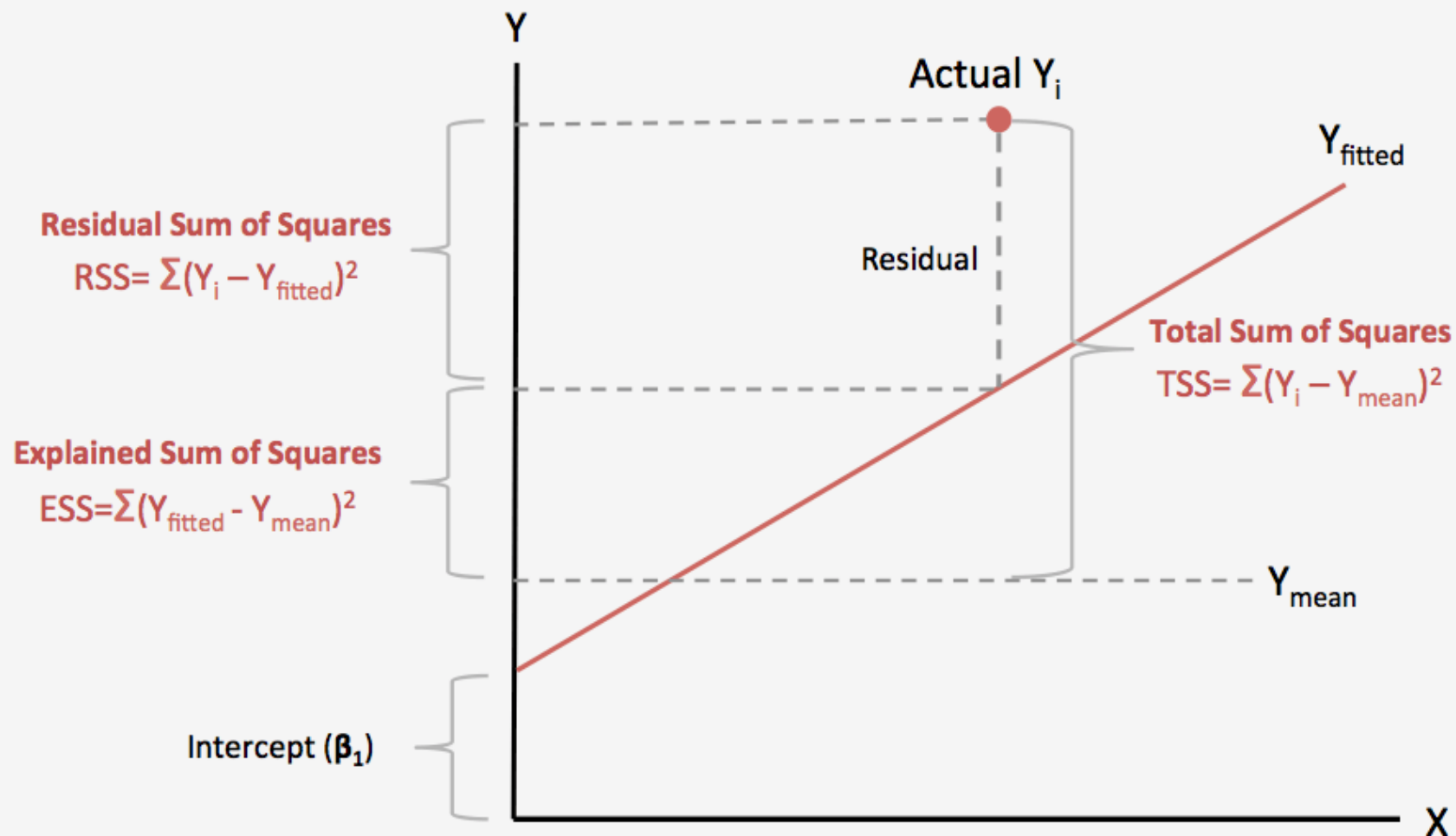
3. 다중회귀분석의 분석 방법

```
# ⑤ 'interactions' 패키지를 이용한 다중회귀분석의 결과표 작성  
library("interactions")  
interact_plot(regression6, pred = "centered.monitor", modx =  
"centered.attachment")
```





R-Squared Explanation



$$R_{Sq} = 1 - \frac{RSS}{TSS}$$

*정규화(Regularization)

1. 전통적인 변수선택 (a.k.a. feature selection) 방법인 stepwise regression은 변수가 많지 않을 때 잘 작동하며, 변수가 많을 때는 정규화(regularization) 방법이 좋은 대안
2. 손실함수(or SSE)에 shrinkage penalty를 추가하여, 변수의 회귀계수들에 정규화 제약을 가함
 - λ (람다) 파라미터를 조절하여 최적의 값을 산출
3. Ex. Ridge/LASSO/Elastic Net
 - LASSO(least absolute shrinkage and selection operator)

*정규화 기법

L1 Regularization(LASSO)

$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M |W_j|$$

L2 Regularization(Ridge)

$$\text{Cost} = \underbrace{\sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2}_{\text{Loss function}} + \lambda \underbrace{\sum_{j=0}^M W_j^2}_{\text{Regularization Term}}$$

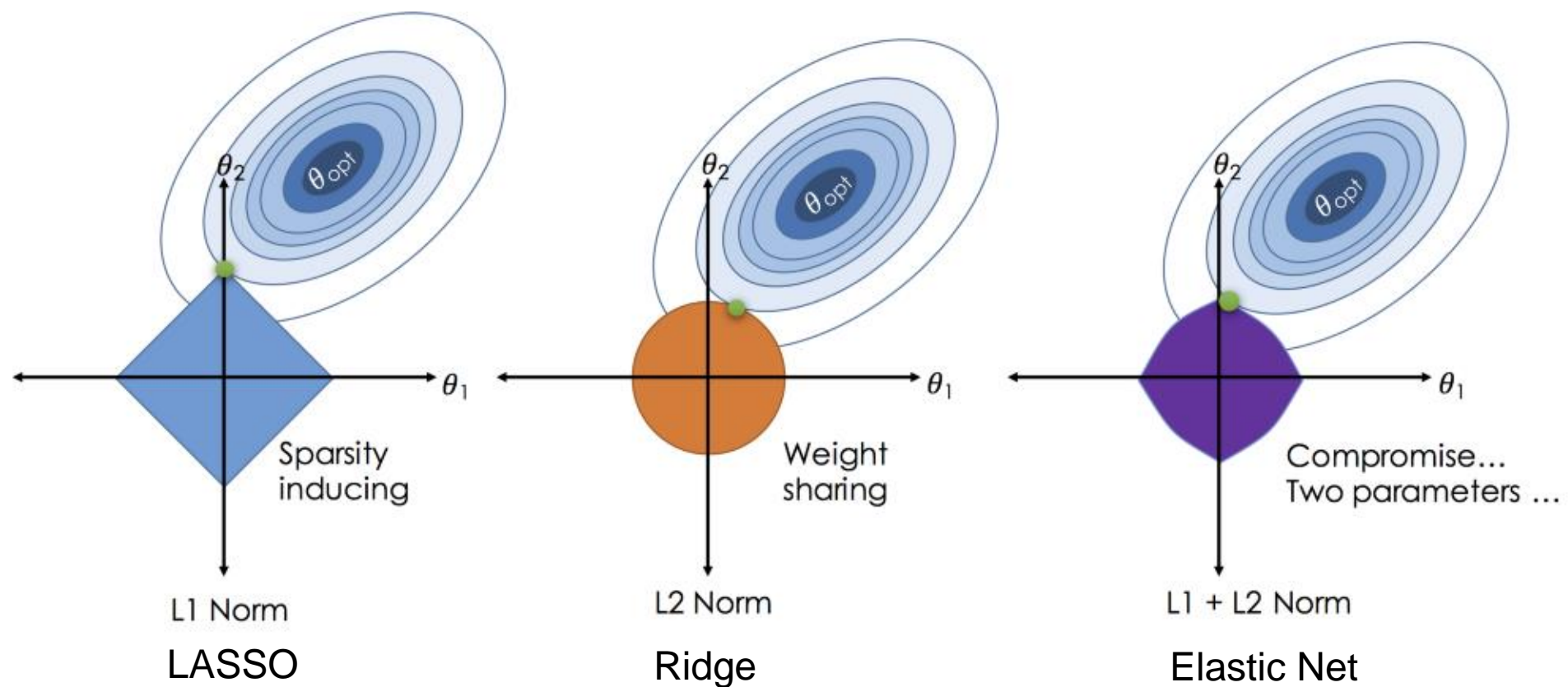
Review of the L_n family of norms

$$L_1 \text{ norm: } \|\mathbf{e}\|_1 = \left[\sum_i |e_i|^1 \right]$$

$$L_2 \text{ norm: } \|\mathbf{e}\|_2 = \left[\sum_i |e_i|^2 \right]^{\frac{1}{2}}$$

$$L_n \text{ norm: } \|\mathbf{e}\|_n = \left[\sum_i |e_i|^n \right]^{\frac{1}{n}}$$

*정규화 기법





*정규화 예제 코드: LASSO/Ridge/Elasticnet

```
##Regularization

library(glmnet)
library(dplyr)

# Remove NA and standardize vars before Lasso/Ridge/ElasticNet
names(spssdata)
reg.var <- spssdata %>% select(c("self.esteem", "sexw1.re",
                                "self.confidence",
                                "attachment", "monitor",
                                "negative.parenting"))
names(reg.var)
str(reg.var)
colSums(is.na(reg.var))
reg.var.valid <- reg.var[complete.cases(reg.var),]
colSums(is.na(reg.var.valid))
reg.var.std <- as.data.frame(scale(reg.var.valid, scale = T))
#standardized matrix plus centering values(mean), scaling
values(sd)
sapply(reg.var.std, sd)
sapply(reg.var.std, mean)
```



```
# Lasso
res.lasso <- glmnet(as.matrix(reg.var.std[2:length(reg.var.std)]),
                    reg.var.std$self.esteem, family = "gaussian",
                    alpha = 1)
res.lasso
#Df: the number of nonzero coefficients for each value of lambda
#Dev: (dev.ratio) the fraction of null deviances explained (for
linear reg., this is R-squared)
plot(res.lasso, xvar = "lambda")
set.seed(12345)
#k-fold cross-validation for glmnet (for lambda parameter tuning)
#cv.glmnet finds the best lambda value
res.lasso <-
cv.glmnet(as.matrix(reg.var.std[2:length(reg.var.std)]),
          reg.var.std$self.esteem,
          family = "gaussian", alpha = 1,
          nfolds = 4, type.measure = "mse")

res.lasso
plot(res.lasso)
log(res.lasso$lambda.min)
log(res.lasso$lambda.1se)
plot(res.lasso$glmnet.fit, xvar = "lambda")
coef.lasso <- coef(res.lasso, s = "lambda.min")[,1]
coef.lasso
mse.min.lasso <- res.lasso$cvm[res.lasso$lambda ==
res.lasso$lambda.min]
#      The mean cross-validated error
mse.min.lasso
r2.min.lasso <- res.lasso$glmnet.fit$dev.ratio[res.lasso$lambda
== res.lasso$lambda.min]
r2.min.lasso
```



```
# Ridge
set.seed(12345)
res.ridge <-
cv.glmnet(as.matrix(reg.var.std[2:length(reg.var.std)]),
           reg.var.std$self.esteem,
           family = "gaussian", alpha = 0,
           nfolds = 4, type.measure = "mse")

res.ridge
plot(res.ridge)
log(res.ridge$lambda.min)
log(res.ridge$lambda.1se)
plot(res.ridge$glmnet.fit, xvar = "lambda")
coef.ridge <- coef(res.ridge, s = "lambda.min")[,1]
coef.ridge
mse.min.ridge <- res.ridge$cvm[res.ridge$lambda ==
res.ridge$lambda.min]
mse.min.ridge
r2.min.ridge <- res.ridge$glmnet.fit$dev.ratio[res.ridge$lambda ==
res.ridge$lambda.min]
r2.min.ridge
```



```
# Elasticnet
set.seed(12345)
res.elastic <-
cv.glmnet(as.matrix(reg.var.std[2:length(reg.var.std)]),
           reg.var.std$self.esteem,
           family = "gaussian", alpha = .5,
           nfolds = 4, type.measure = "mse")

res.elastic
plot(res.elastic)
log(res.elastic$lambda.min)
log(res.elastic$lambda.1se)
plot(res.elastic$glmnet.fit, xvar = "lambda")
coef.elastic <- coef(res.elastic, s = "lambda.min")[,1]
coef.elastic
mse.min.elastic <- res.elastic$cvm[res.elastic$lambda ==
res.elastic$lambda.min]
mse.min.elastic
r2.min.elastic <-
res.elastic$glmnet.fit$dev.ratio[res.elastic$lambda ==
res.elastic$lambda.min]
r2.min.elastic

# Compare with lm output
res.lm <- lm(self.esteem ~ sexw1.re+self.confidence+
              attachment+monitor+negative.parenting, data =
reg.var.std)
summary(res.lm)
library(dvMisc)
get_mse(res.lm)
```

USING R *회귀모델 성능평가지표

Mean squared error

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

Mean absolute error

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

Mean absolute percentage error

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right|$$
