



통계실습 I. 회귀분석, ARIMA 모형

가천대 길병원 고급통계교육

방태모

2022-06-14

목차

1 회귀분석

- 방법론 소개
- 단순회귀모형
- 실습: 단순회귀모형
- 다중회귀모형
- 실습: 다중회귀모형

2 ARIMA 모형: 실제 사례 중심

- 준비
- 결측치 처리
- 시계열 자료 다루기
- 모형 적합
- 예측

1 회귀분석

방법론 소개

- 데이터로부터 독립변수(또는 예측변수, x)들의 함수를 다음과 같이 추정하여 종속변수(또는 반응변수, y)를 예측하는 방법:

$$\hat{y} = f(x_1, x_2, \dots, x_p). \quad (1)$$

- 여기서 $x = (x_1, x_2, \dots, x_p)^T$ 는 주어지는 값
- y 는 주어진 x 에서 측정되는 값으로 벡터로 주어질 수 있음

단순회귀모형

- 단순선형회귀(simple linear regression)는 예측변수와 종속변수가 모두 1개인 경우에 해당:

$$y = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim N(0, \sigma^2). \quad (2)$$

- 오차항 $\epsilon_i \sim i.i.d N(0, \sigma^2)$ 에 관한 가정
 - 정규성(Normality)
 - 독립성(Independent): $\text{Cov}(\epsilon_i, \epsilon_j) = 0(i \neq j)$
 - 등분산성(Homoscedasticity): $\text{Var}(\epsilon_i) = \sigma^2$

단순회귀모형

Observation Number	Temperature (x_i)	Yield (y_i)
1	50	122
2	53	118
3	54	128
4	55	121
5	56	125
6	59	136
7	62	144
8	65	142
9	67	149
10	71	161
11	72	167
12	74	168
13	75	162
14	76	171
15	79	175
16	80	182
17	82	180
18	85	183
19	87	188
20	90	200
21	93	194
22	94	206
23	95	207
24	97	210
25	100	219

fig 1. Data

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (3)$$

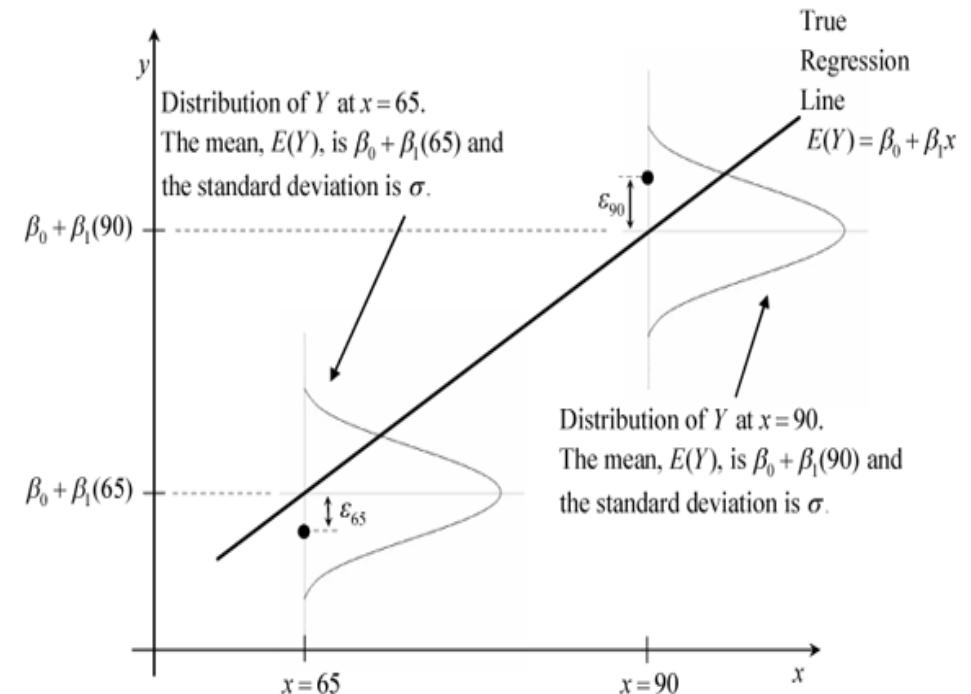


fig 2. Simple linear regression

실습: 단순회귀모형

준비하기

- 패키지 설치

```
install.packages("MASS")
install.packages("car")
install.packages("dplyr")
```

- 패키지 로딩

```
library(MASS)
library(car)
library(dplyr)
```

실습: 단순회귀모형

데이터 소개

- 고양이들의 성별, 몸무게(Bwt), 심장몸무게(Hwt)에 관한 자료 `cats` 이용

```
head(cats, 4)
```

Sex	Bwt	Hwt
F	2.0	7.0
F	2.0	7.4
F	2.0	9.5
F	2.1	7.2

- 데이터 구조 확인

```
glimpse(cats)
```

```
## Rows: 144  
## Columns: 3  
## $ Sex <fct> F,  
## $ Bwt <dbl> 2.0, 2.0, 2.0, 2.1, 2.1, 2.1,  
## $ Hwt <dbl> 7.0, 7.4, 9.5, 7.2, 7.3, 7.6
```

실습: 단순회귀모형

모형 적합

- {stats} 패키지의 `lm()`을 이용해 단순회귀모형 적합

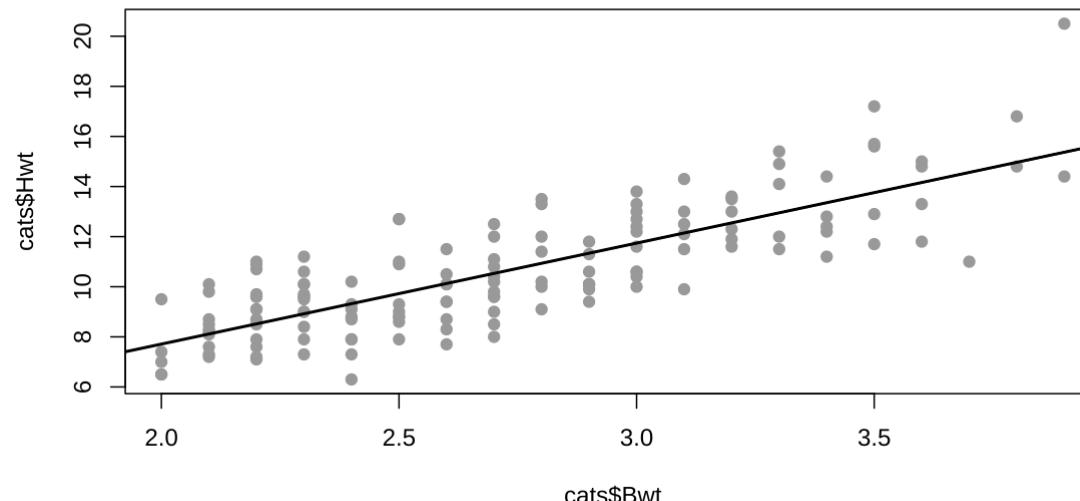
```
lm_mod <- lm(Hwt ~ Bwt, data = cats)  
summary(lm_mod)
```

```
##  
## Call:  
## lm(formula = Hwt ~ Bwt, data = cats)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -3.5694 -0.9634 -0.0921  1.0426  5.1238  
##  
## Coefficients:
```

실습: 단순회귀모형

회귀선 시각화

```
plot(cats$Hwt ~ cats$Bwt, pch = 19, col = "darkgray")
abline(lm_mod, lwd = 2)
```

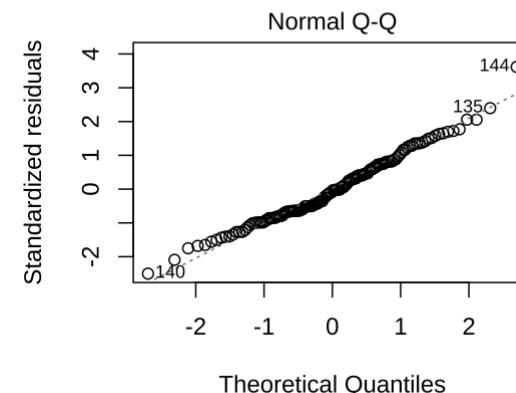
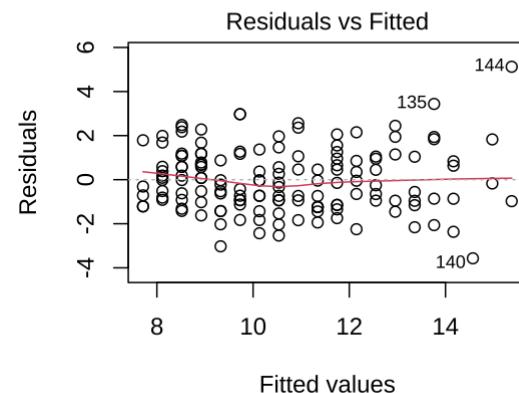


실습: 단순회귀모형

잔차분석: 등분산성, 정규성

- y 와 x 간의 relationship에 관한 분석을 목적으로 선형회귀모형을 적합할 때는 잔차분석 필수

```
par(mfrow = c(1, 2))
plot(lm_mod, 1)
plot(lm_mod, 2)
```



실습: 단순회귀모형

잔차분석: 이상점 체크

- 144번 자료에 대한 정보 확인

```
cats[144, ]
```

```
##      Sex Bwt Hwt
## 144     M  3.9 20.5
```

```
lm_mod$fitted[144]
```

```
##          144
## 15.37618
```

```
lm_mod$residuals[144]
```

```
##          144
## 5.123818
```

실습: 단순회귀모형

잔차분석: 독립성 검정

- Durbin-Watson Test를 이용한 잔차의 독립성 검정
 - 더빈-왓슨 테스트 통계량(d)는 $0 < d < 4$ 사이의 값을 가짐
 - $d \rightarrow 2$: 자기상관이 존재하지 않음
 - $d \rightarrow 0$ or $d \rightarrow 4$: 자기상관이 존재함
- {car} 패키지의 `durbinWatsonTest()`를 이용해 잔차에 대해 독립성 검정 수행

```
residuals(lm_mod) |>  
durbinWatsonTest()
```

```
## [1] 1.579896
```

다중회귀모형

- 다중선형회귀(multiple linear regression)는 단순선형회귀의 확장으로 독립변수의 수가 여러 개인 경우에 해당:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \epsilon \sim N(0, \sigma^2) \quad (4)$$

- 오차항에 대한 가정은 단순선형회귀의 경우와 같음

다중회귀모형

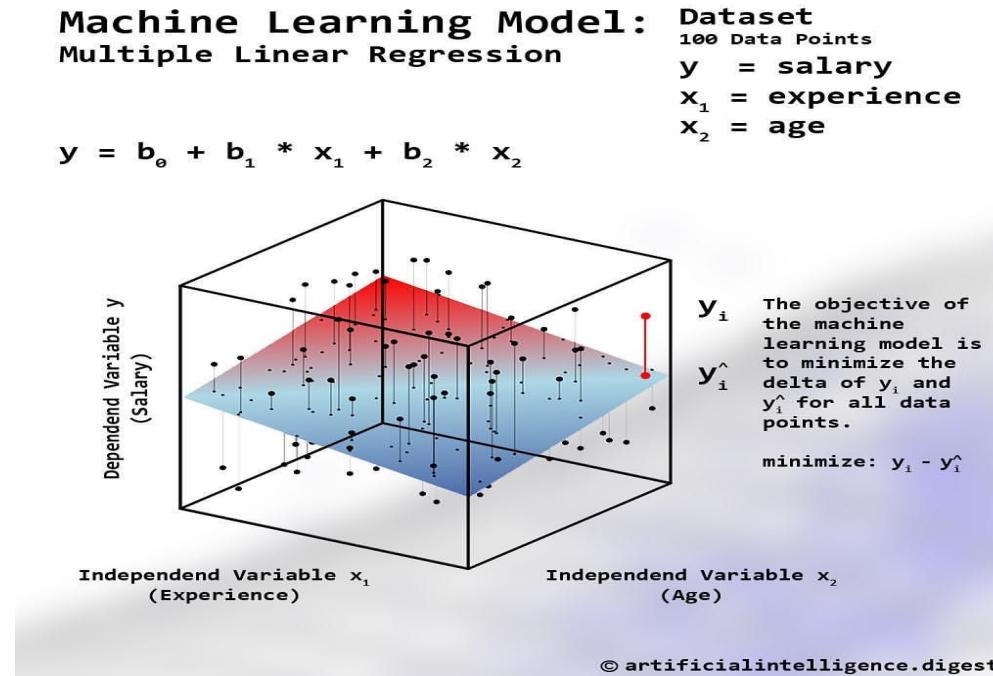


fig 3. Multilple linear regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad (5)$$

다중회귀모형

$$\boldsymbol{\beta} = (X^T X)^{-1} X^T \mathbf{y} \quad (6)$$

- 다중공선성 문제

- 다중선행회귀에서는 자료의 수(n)가 모수의 수(p)보다 필히 많아야 함($n \geq p + 1$)
- $p \geq n$ 인 경우 또는 다중공선성이 심한 경우 ($X^T X$)의 역행렬의 계산이 불가능해짐
- 결과적으로 $\boldsymbol{\beta}$ 추정량의 분산이 매우 커짐
- 즉, 추정된 회귀계수를 신뢰하기 어려움

- 다중공선성의 징후

- X 들 간에 높은 상관계수가 존재할 때
- 이론적으로 반응변수와 상관이 높을것으로 생각되는 변수임에도 회귀계수가 유의하지 않음
- 이론적으로 X 에 따라 반응변수가 증가해야 함에도 회귀계수가 음의 값을 가짐
- 특정 변수 X 를 추가하거나 제외하였을 때, 회귀계수의 변화가 크게 일어남

다중회귀모형

- 다중공선성 문제 해결 방법
 - 상관된 예측변수 제거: 상관성이 높은 변수를 제거해 나가되, R^2 가 높은 모형 선택
 - Variable selection: Forward Selection, Backward elimination, Stepwise selection
 - Best subset regression
 - 벌점회귀(Penalized regression) 이용
 - Lasso regression

실습: 다중회귀모형

데이터 소개

- 미국 50개 주에서 여러 변수값(인구, 수입, 문맹비율, 기대수명, 살인율, 고졸비율, 연평균영하기 온일수, 면적)을 측정한 자료
- 이 중 기대수명(Life exp)을 반응변수로 다중회귀분석 실시

```
head(state.x77, 3)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417

실습: 다중회귀모형

데이터 소개

- 데이터 구조 확인

```
glimpse(state.x77)
```

```
##  num [1:50, 1:8] 3615 365 2212 2110 21198 ...
##  - attr(*, "dimnames")=List of 2
##    ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##    ..$ : chr [1:8] "Population" "Income" "Illiteracy" "Life Exp" ...
```

- 데이터프레임으로 변환 후, 변수명 비카 제외

실습: 다중회귀모형

모형 적합

```
lm_mod2 <- lm(life_exp ~ ., data = state)  
summary(lm_mod2)
```

```
##  
## Call:  
## lm(formula = life_exp ~ ., data = state)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -1.48895 -0.51232 -0.02747  0.57002  1.49447  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)
```

실습: 다중회귀모형

다중공선성 검토: 분산팽창요인(VIF, Variance Inflation factor)

$$VIF_j = \frac{1}{1 - R_j^2} \quad (7)$$

- R_j^2 는 j -번째 예측변수(x_j)를 나머지($p - 1$)개의 예측변수로 회귀를 수행하여 구해지는 결정계수
- 10보다 큰 경우 다중공선성을 야기하는 변수로 간주

- {car} 패키지의 `vif()`를 이용해 구할 수 있음

```
vif(lm_mod2)
```

```
## population      income illiteracy      mur
##   1.499915    1.992680    4.403151    2.616
```

실습: 다중회귀모형

- Variable selection: Backward elimination
 - `step()` 이용

```
lm_mod3 <- step(lm_mod2, direction = "backward", trace = FALSE)
summary(lm_mod3)
```

```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##      data = state)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
```

실습: 다중회귀모형

- 신뢰구간 구하기

```
confint(lm_mod3)
```

```
##                   2.5 %      97.5 %
## (Intercept) 6.910798e+01 72.9462729104
## population -4.543308e-07  0.0001007343
## murder      -3.738840e-01 -0.2264135705
## hs_grad      1.671901e-02  0.0764454870
## frost       -1.081918e-02 -0.0010673977
```

- 주어진 자료에 대한 예측값 구하기

```
predict(lm_mod3, list(population = 4000, murder = 10.5, hs_grad = 48, frost = 100))
```

2 ARIMA 모형: 실제 사례 중심

준비

- 패키지 설치

```
install.packages("fpp3")
install.packages("readr")
install.packages("showtext")
install.packages("ggplot2")
install.packages("conflicted")
```

- 패키지 불러오기

```
library(fpp3)
library(readr)
library(showtext)
library(ggplot2)
library(ggh4x)
```

준비

- 2010년 1월 - 2021년 9월 월별 익명의 질환 발생 건수에 관한 자료

```
disease <- read_csv("./data/example.csv")
disease
```

```
## # A tibble: 1,128 × 4
##   Sex     Group    Date      N
##   <chr>   <chr>    <chr>    <dbl>
## 1 Female  10_14세 2010 Jan    114
## 2 Female  10_14세 2010 Feb    121
## 3 Female  10_14세 2010 Mar    106
## 4 Female  10_14세 2010 Apr    120
## 5 Female  10_14세 2010 May    112
## 6 Female  10_14세 2010 Jun    122
## 7 Female  10_14세 2010 Jul    116
```

- 8개 그룹에 대해 모델링 필요

```
disease |>
  select(-Date, -N) |>
  distinct()
```

```
## # A tibble: 8 × 2
##   Sex     Group
##   <chr>   <chr>
## 1 Female  10_14세
## 2 Female  15_19세
## 3 Female  5_9세
## 4 Female  5세미만
## 5 Male    10_14세
## 6 Male    15_19세
## 7 Male    5_9세
```

결측치 처리

- 결측이 존재하지 않는 데이터이지만 임의로 결측 처리(원자료 값 121)

```
disease2 <- disease |>  
  slice(-2)  
disease2
```

```
## # A tibble: 1,127 × 4  
##   Sex    Group Date      N  
##   <chr>   <chr>  <chr>    <dbl>  
## 1 Female 10_14세 2010 Jan     114  
## 2 Female 10_14세 2010 Mar     106  
## 3 Female 10_14세 2010 Apr     120  
## 4 Female 10_14세 2010 May     112  
## 5 Female 10_14세 2010 Jun     122  
## 6 Female 10_14세 2010 Jul     116  
## 7 Female 10_14세 2010 Aug     138  
## 8 Female 10_14세 2010 Sep     122
```

시계열 자료 다루기

- R에게 우리가 모델링할 시계열이 총 8개임을 알려줘야함
- `as_tsibble()`을 이용할 수 있음

```
disease3 <- disease2 |>
  mutate(Date = yearmonth(Date)) |>
  as_tsibble(key = c(Sex, Group), index = Date)
disease3
```

```
## # A tsibble: 1,127 x 4 [1M]
## # Key:      Sex, Group [8]
##   Sex     Group       Date     N
##   <chr>   <chr>     <mth> <dbl>
## 1 Female  10_14세 2010 Jan    114
## 2 Female  10_14세 2010 Mar    106
## 3 Female  10_14세 2010 Apr    120
```

시계열 자료 다루기

- `fill_gaps()` 이용 결측치 대치

```
disease4 <- disease3 |>  
  fill_gaps(N = 121L, .full = TRUE)  
disease4
```

```
## # A tsibble: 1,128 x 4 [1M]  
## # Key:      Sex, Group [8]  
##   Sex     Group     Date     N  
##   <chr>   <chr>     <mth> <dbl>  
## 1 Female  10_14세 2010 Jan    114  
## 2 Female  10_14세 2010 Feb    121  
## 3 Female  10_14세 2010 Mar    106  
## 4 Female  10_14세 2010 Apr    120  
## 5 Female  10_14세 2010 May    112  
## 6 Female  10_14세 2010 Jun    122
```

모형 적합

- AICc를 기반으로 최적의 ARIMA 모형 적합

```
mod <- disease4 |>  
  model(arima = ARIMA(N))  
mod
```

```
## # A mable: 8 x 3  
## # Key:      Sex, Group [8]  
##   Sex     Group           arima  
##   <chr>   <chr>           <model>  
## 1 Female  10_14세 <ARIMA(0,1,1)>  
## 2 Female  15_19세 <ARIMA(1,1,1)>  
## 3 Female  5_9세    <ARIMA(0,1,1)>  
## 4 Female  5세미만 <ARIMA(0,1,1)>  
## 5 Male    10_14세 <ARIMA(0,1,1)>  
## 6 Male    15_19세 <ARIMA(0,1,1)>
```

예측

- 향후 15개월 예측 후, 80% 및 95% 신뢰구간 추출

```
f <- mod %>%
  forecast(h = "15 months") %>%
  hilo(level = c(80, 95)) %>%
  unpack_hilo(c("80%", "95%")) # 신뢰구간 값 열로 추출
head(f, 3)
```

Sex	Group	.model	Date	N	.mean	80%_lower	80%_upper	95%_lower	95%_upper
Female	10_14 세	arima	2021 Oct	N(99, 133)	99.20714	84.43298	113.9813	76.61202	140.8534
Female	10_14 세	arima	2021 Nov	N(99, 133)	99.20714	84.40289	114.0114	76.56600	140.7385
Female	10_14 세	arima	2021 Dec	N(99, 134)	99.20714	84.37286	114.0414	76.52007	140.7385

시각화

- 과거 시점(2010년 1월 - 2021년 9월)의 자료 만들기

```
historic <- bind_rows(  
  disease4 %>%  
    mutate(Type = "과거 실제값") %>%  
    as_tibble(),  
  mod %>%  
    fitted %>%  
    mutate(Type = "모형 적합값") %>%  
    as_tibble %>%  
    rename(N = .fitted) %>%  
    select(Date, Sex, Group, N, Type)  
)
```

시각화

- 미래 시점(2021년 10월 - 2022년 12월)의 자료 만들기

```
fore <- f %>%
  mutate(Type = "모형 예측값", N = .mean) %>%
  as_tibble %>%
  select(Date, Sex, Group, N, Type)
```

시각화

- `{ggplot2}` 이용

```
ggplot() +  
  geom_line(data = historic,  
            aes(x = Date, y = N, col = Type)) + # 과거 실제값, 모형 적합값  
  geom_line(data = fore,  
            aes(x = Date, y = N, col = Type)) + # 모형 예측값  
  geom_ribbon(data = f, # 80% 신뢰구간  
              aes(x = Date, ymin = `80%_lower`, ymax = `80%_upper`),  
              fill = "skyblue", alpha = 0.25) +  
  geom_ribbon(data = f, # 95% 신뢰구간  
              aes(x = Date, ymin = `95%_lower`, ymax = `95%_upper`),  
              fill = "skyblue", alpha = 0.25/2) +  
  scale_color_manual(values = c("tomato", "blue", "paleturquoise3"), name = "") +  
  theme(  
    text = element_text(family = "nanum", size = 12)),
```

시각화

