

2-2 기초 통계 분석

제2장 통계분석

제1절 통계학 개론

제2절 기초 통계 분석

기술통계(Descriptive Statistics)

자료의 특성을 표, 그림, 통계량 등을 사용하여 쉽게 파악할 수 있도록 정리/요약하는 것

- 자료를 요약하는 기초적 통계
- 데이터 분석에 앞서 데이터의 대략적인 통계적 수치를 계산함으로써 데이터에 대한 대략적인 이해와 앞으로 분석에 대한 통찰력을 얻기에 유리

회귀 분석(Regression Analysis)

1. 회귀분석의 개요

- **정의:** 하나 이상의 독립변수가 종속변수에 미치는 영향을 추정하는 통계 기법.
- **목적:** 변수들 간 인과관계를 밝히고 모형을 적합시켜 예측 및 추론.
- **종류**
 - 단순 선형 회귀: 독립변수 1개
 - 다중 선형 회귀: 독립변수 2개 이상
- **변수 구분**
 - y: 종속변수, 반응변수
 - x: 독립변수, 설명변수
 - 영향을 받는 변수(y) : 반응변수(Response Variable), 종속변수(Dependent Variable), 결과변수(Outcome Variable)

- 영향을 주는 변수(x) : 설명변수(Explanatory Variable) , 독립변수(Independent Variable), 예측 변수(Predictor Variable)

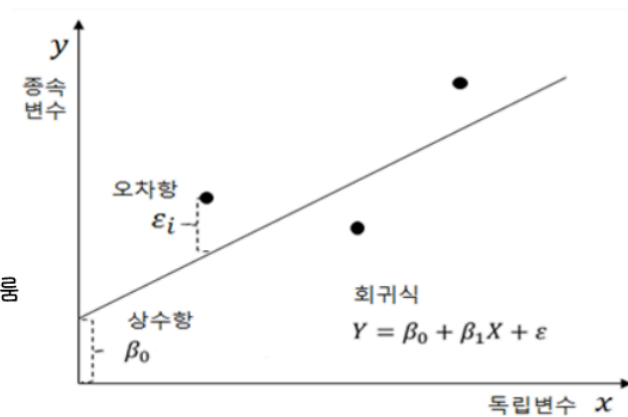
단순선형회귀분석

하나의 독립변수가 종속변수에 미치는 영향을 추정하는 통계기법

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon, \quad i = 1, 2, \dots, n, \quad \varepsilon_i \sim N(0, \sigma^2)$$

iid(독립적이고 동일한 분포)
-independent
-identically
-distributed

- $-y_i$: i번째 종속변수 값
- $-x_i$: i번째 독립변수 값
- $-\beta_0$: 선형 회귀식의 절편
- $-\beta_1$: 선형 회귀식의 기울기
- $-\varepsilon$: 오차항. 독립적이며 $N(0, \sigma^2)$ 의 분포를 이룸



출처: <https://velog.io/@dangdang/ADsP-통계-분석>

회귀분석에서의 검토사항

- 회귀계수들이 유의미한가?
: 해당 계수의 t-통계량(평균)의 p-값이 0.05보다 작으면 회귀계수가 통계적으로 유의
- 모형이 얼마나 설명력을 갖는가?
: 결정계수(R^2) 확인
-결정계수는 0~1값을 가지며, 높은 값을 가질 수록 추정된 회귀식의 설명력이 높음
- 모형이 데이터를 잘 적합하고 있는가?
: 잔차를 그래프로 그리고 회귀진단

선형회귀분석의 가정

- 선형성
: 입력변수와 출력변수의 관계가 선형(선형회귀분석에서 가장 중요한 가정)
- 등분산성(분산이 같음)
: 오차의 분산이 입력변수와 무관하게 일정
 - 잔차플롯(산점도)을 활용하여 잔차(표본으로 추정된 회귀식과 실제 관측값의 차이)와 입력변수간에 아무런 관련성이 없게 무작위적으로 고루 분포되어야 등분산성 가정 만족
- 독립성
: 입력변수와 오차는 관련 무관
 - 자기상관(독립성)을 알아보기 위해 Durbin-Waston 통계량 사용
 - 시계열 데이터에서 많이 활용
- 비상관성
: 오차들끼리 무 상관
- 정상성(정규성)
: 오차의 분포가 정규분포를 따름
 - Q-Q plot, Kolmogorov-Smirnov 검정, Shapiro-Wilk 검정등을 활용해 정규성 확인

회귀분석의 종류 비교표

종류	모형	설명
단순회귀	$Y = \beta_0 + \beta_1 X + \epsilon$	설명변수가 1개이며, 반응변수와의 관계가 직선
다중회귀	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$	설명변수가 k 개이며, 반응변수와의 관계가 선형 (1차 함수)
다항회귀	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$ (예: $k = 2$ 일 때)	설명변수가 k 개이며, 반응변수와의 관계가 1차 이상 함수 (비선형 포함 가능)
비선형회귀	$Y = g(\beta_0 + \beta_1 X + \dots + \beta_k X_k) + \epsilon$	회귀식의 형태가 미지의 계수 β_i 들이 선형관계를 이루지 않는 경우. 예: $g(t) = e^t$ 형태의 지수함수 등

회귀 계수 검정

결정계수 (R^2)

- $SSR/SST, 0 \leq R^2 \leq 1$
- 단순회귀에서는 상관계수 r 의 제곱

회귀 직선의 적합도 검토

- 결정계수 R^2 통해 회귀 식의 타당성 검토
- 수정된 결정계수 (Adjusted R^2): 독립변수 수 증가 시 보정

회귀분석에서의 검토사항

1. 회귀 계수들이 유의미한가?
 - t 통계량의 p값이 0.05보다 작으면 유의미함
2. 모형은 얼마나 설명력을 갖는가?
 - 결정계수(R^2) 확인: Multiple R-squared:
 - 수정된 결정계수(R^2) 확인: Adjusted R-squared:
3. 모형이 데이터를 잘 적합하고 있는가?
 - 잔차를 그래프로 그리고 회귀 진단

회귀계수의 추정 (최소 제곱 법, 최소 자승 법)

- 잔차 제곱이 가장 작은 선을 구하는 방법

회귀계수의 t-value 해석

기본 수식

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- β_j : 설명변수 X_j 에 대한 회귀계수의 추정값
- $SE(\beta_j)$: 해당 회귀계수의 표준오차(Standard Error)
 - 계수의 불확실성을 수치로 표현한 것

t-value가 크다는 것은?

- t-value 가 클수록, 해당 계수는 0이 아닐 가능성이 높다는 의미야.
 - 회귀 계수가 0이라는 귀무가설($H_0: \beta = 0$)을 기각할 수 있게 돼.
- 일반적으로는 다음 기준을 참고해:
 - $|t| > 2 \rightarrow$ 통계적으로 유의할 가능성 있음
 - $|t| > 3 \sim 4$ 이상이면 거의 확실하게 유의미
 - $|t| \gg 10$ 이면 아주 강력한 신호

회귀분석의 검정

1. 회귀계수의 검정

- 귀무가설: $\beta_1 = 0$ (x와 y는 무관)
- 대립가설: $\beta_1 \neq 0$

2. 결정계수

- $SST = SSR + SSE$
- 전체 제곱합, 회귀 제곱합, 오차 제곱합 구분

분산 분해의 3요소

1. SST (Total Sum of Squares)

- 전체 데이터의 y 값이 평균으로부터 얼마나 퍼져있는지 측정
- 수식: $SST = \sum (y_i - \bar{y})^2$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

- “총 분산의 양”이라고 보면 돼
- 아무런 모델 없이 \bar{y} 만 가지고 예측할 때 생기는 총 오차

2. SSR (Regression Sum of Squares)

- 모델이 y 값을 예측할 때 평균보다 얼마나 더 잘 예측했는지 나타냄
- 수식: $SSR = \sum (\hat{y}_i - \bar{y})^2$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- 즉, 예측값이 평균값보다 얼마나 더 정확한지를 측정
- 모델이 설명한 분산 = 설명된 분산

3. SSE (Error Sum of Squares)

- 실제 y 값과 예측값 사이의 **오차(잔차)**를 측정함
- 수식: $SSE = \sum (y_i - \hat{y}_i)^2$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- 모델이 설명하지 못한 나머지 분산 = 잔차 분산

SST = SSR + SSE 공식 이해하기

이 식은 다음과 같은 의미야:

전체 y의 총 변동성(SST)은

모델이 설명한 부분(SSR) + 설명 못한 오차(SSE)로 나뉜다.

즉, 이 회귀모델이 얼마나 잘 맞았는지를 **설명된 분산 대 전체 분산 비율로 측정**할 수 있다는 거야.

제곱합 총계(SST):

제곱합은 종속 변수 Y의 참값과 평균값의 차이 제곱의 합입니다. 이는 현재 데이터 집합에서 모형으로 설명할 수 있는 총 변동성을 나타냅니다.

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

몇 가지 주요 지침:

- 총 변동성 **(SST)** = **(SSR)**설명 변동성 + **(SSE)**설명되지 않은 변동성
- 총 제곱합(TSS)

으로도 알려져 있습니다 .

제곱 오차의 합(SSE):

제곱 오차의 합은 회귀 모형의 추정력을 나타냅니다. 이는 학습 후 모형으로 설명할 수 없었던 분산의 양을 나타냅니다.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

몇 가지 주요 지침:

- 높은 추정력을 가진 모델은 예측값(\hat{Y})이 실제값(Y)에 가깝다는 것을 의미합니다. 따라서 최종 목표인 SSE 값이 낮아집니다.

- 실제 값과 예측 값의 차이가 제공되므로 차이가 클 경우 큰 페널티가 부과됩니다.
- 잔차 제곱합(RSS)
으로도 알려져 있습니다 .

제곱합 회귀(SSR):

제곱합 회귀는 모델이 데이터 포인트에 얼마나 잘 적합한지를 설명합니다. 이는 모델이 훈련 후 설명할 수 있었던 분산의 양을 나타냅니다.

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

회귀 모형의 적합도를 계산하는 데 사용되는 R-제곱 지표는 설명 변동성(SSR) 과 총 변동성(SST) 의 비율에 불과합니다 .

$$R\text{-제곱} = SSR/SST$$

범위는 $[0,1]$ 입니다.

- **SSR이 SST에 가까워질수록 R- 제곱 값은 1에 가까워집니다.**
 - 이는 모델이 적합도가 높고 데이터의 많은 변동성을 설명할 수 있음을 의미합니다.

몇 가지 주요 지침:

- **ESS(설명된 제곱합)** 로도 알려짐

Formula: $\hat{Y} = b + w * X$

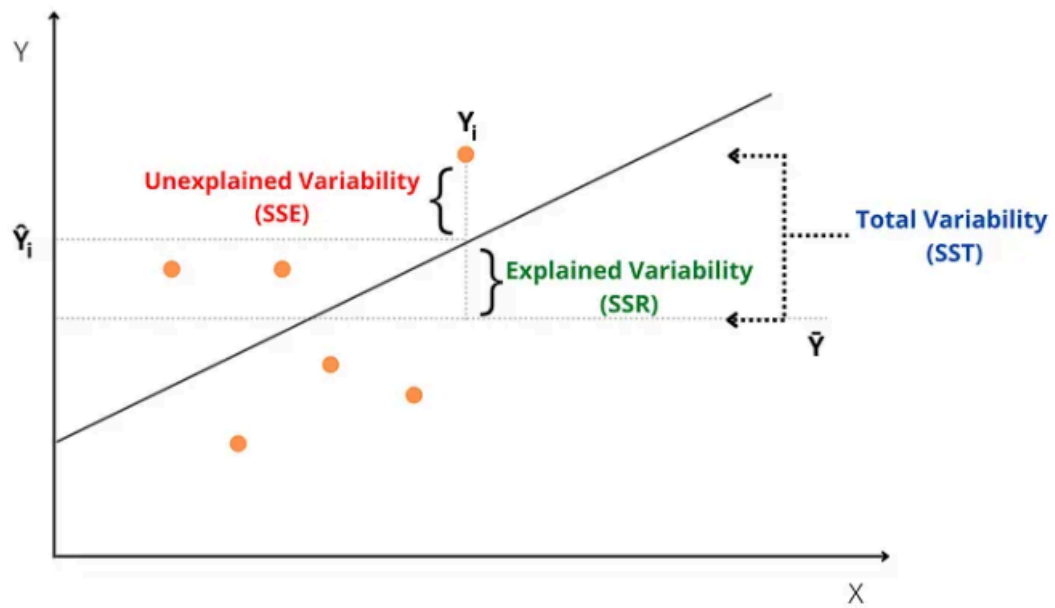
\hat{Y} : Predicted value of Y

b : Y -intercept (bias factor)

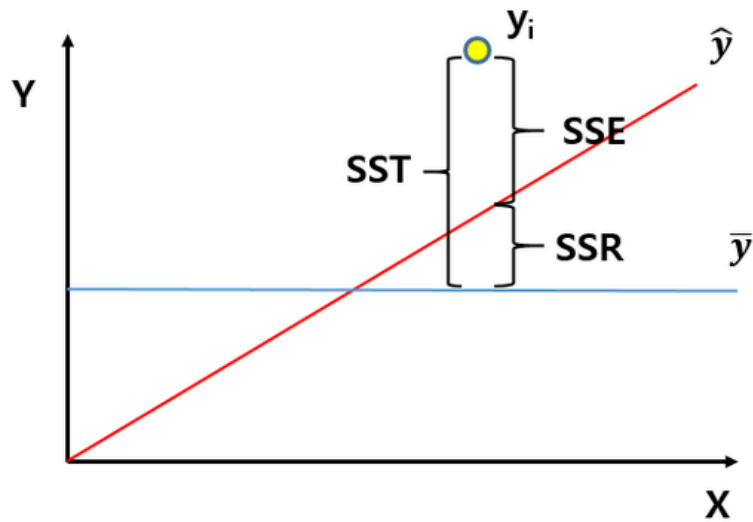
w : Weights

X : Independent Variable X

\bar{Y} : Mean of True values of Y



SST, SSE & SSR



SST(Y의 전체 변동) : $\sum (y_i - \bar{y})^2$
 SSR(모형에 의해 설명되는 변동) : $\sum (\hat{y}_i - \bar{y})^2$
 SSE(모형에 의해 설명이 되지 않는 변동) : $\sum (y_i - \hat{y}_i)^2$

이름	수식	의미
SST	$\sum (y_i - \bar{y})^2$	전체 변동성
SSR	$\sum (\hat{y}_i - \bar{y})^2$	모델이 설명한 변동성
SSE	$\sum (y_i - \hat{y}_i)^2$	모델이 설명 못한 오차

다중 선형 회귀분석

- 회귀식:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

- F 통계량 검정:

유의수준 5%에서 p값이 0.05보다 작으면 모형 유의

최적회귀방정식: 최적회귀방정식의 선택

설명변수(x) 선택: 필요한 변수만 상황에 따라 타협을 통해 선택

- y에 영향을 미칠 수 있는 모든 설명변수 x들을 y의 값을 예측하는 데 참여
- 데이터에 설명변수 x들의 수가 많아지면 관리하는데 많은 노력이 요구되므로
 - 가능한 범위 내에서 적은 수의 설명변수 포함

모형선택(Exploratory Analysis): 분석 데이터에 가장 잘 맞는 모형을 찾아내는 방법

- 모든 가능한 조합의 회귀분석(All Possible Regression) : 모든 가능한 독립변수들의 조합에 대한 회귀모형을 생성한 뒤 가장 적합한 회귀모형 선택

단계적 변수선택(Stepwise Variable Selection)

- 전진선택법(Forward Selection)
 - 절편만 있는 상수모형으로부터 시작해 중요하다고 생각되는 설명변수부터 차례로 모형에 추가
 - 변수의 개수가 많은 경우에도 사용 가능
 - 변수값의 작은 변동에도 그 결과가 크게 달라짐 =>안정성 부족
- 후진제거법(Backward Elimination)
 - 독립변수 후보 모두를 포함한 모형에서 출발해 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때의 모형 선택
 - 변수의 개수가 많은 경우 사용하기 어려움
 - 전체 변수들의 정보를 이용
- 단계선택법(Stepwise Method)
 - 전진선택법에 의해 변수를 추가하면서 새롭게 추가된 변수에 기인해 기존 변수의 중요도가 약화되면 해당 변수를 제거하는 등 단계별로 추가 또는 제거되는 변수의 여부를 검토해 더 이상 없을 때 중단

최적 회귀 방식

변수 선택법

방법	설명
전진선택법	절편만으로 시작, 변수 추가
후진제거법	모든 변수 포함 후 하나씩 제거
단계선택법	추가/제거 병행

모형 선택 기준: AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion)

- 벌점화된 선택기준: 모형의 복잡도에 벌점을 주는 방법
 - 모든 후보 모형들에 대해 AIC 또는 BIC를 계산하고, 그 값이 최소가 되는 모형을 선택
 - 모형 선택의 일치성(Consistency Inselection): 자료의 수가 늘어날 때 참인 모형이 주어진 모형 선택 기준의 최소값을 갖게 되는 성질

실습

MASS::Animals 데이터셋 설명

1. 데이터 개요

- 패키지: `MASS` (Modern Applied Statistics with S)
- 데이터명: `Animals`
- 행(Row): 28개 (동물의 종류)
- 열(Column): 2개
 - `body`: 몸무게 (Body weight in kg)
 - `brain`: 뇌 무게 (Brain weight in g)
- 각 행의 이름이 동물 이름으로 되어 있음 (예: "Mouse", "Elephant", "Cat" 등)

코드

```
m <- lm(y ~ u+v+w)
summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min       1Q   Median       3Q      Max
-0.188562 -0.058632 -0.002013  0.080024  0.143757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.041653   0.264808  11.486 2.62e-05 ***
u             0.123173   0.012841   9.592 7.34e-05 ***
v             1.989017   0.016586  119.923 2.27e-11 ***
w            -2.997816   0.005421 -552.981 2.36e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1303 on 6 degrees of freedom
Multiple R-squared:  1, Adjusted R-squared:  1
F-statistic: 1.038e+05 on 3 and 6 DF, p-value: 1.564e-14
```

✓ 요약

- 이 회귀모델은 종속변수 `y`를 설명하기 위해 독립변수 `u`, `v`, `w`를 사용함.
- R^2 값이 거의 1이고, 모든 변수의 **p-value가 매우 작아** 통계적으로 유의미함.
- 잔차도 작고 F-검정 결과도 매우 강력함 → 모델 설명력이 극도로 높음.

선형회귀 결과 요약 설명

1. 모델 개요

- 모델 수식: $y = \beta_0 + \beta_1 u + \beta_2 v + \beta_3 w + \epsilon$

$$y = \beta_0 + \beta_1 u + \beta_2 v + \beta_3 w + \epsilon$$

- 호출 내용: `lm(formula = y ~ u + v + w)`
→ y를 u, v, w로 설명하는 선형모델을 만든 거임

2. 잔차 분석 (Residuals)

- 최소값 ~ 최대값: 0.188562 ~ 0.143757
- 중앙값(Median): 0.002013
- 대부분 잔차가 ±0.1 정도로 매우 작음
→ 예측값이 실제값과 거의 일치함

- 잔차란 관측값과 예측값의 차이

$$e_i = y_i - \hat{y}_i$$

- 여기서:
 - y_i : 실제 관측값
 - \hat{y}_i : 예측된 값 (회귀선 위의 값)
 - e_i : i번째 관측값의 잔차

3. 회귀계수(Coefficients)

변수	계수 (Estimate)	표준오차	t값	p-value	해석
(Intercept)	3.041653	0.264808	11.486	2.62e-05	절편, y축 시작값
u	0.123173	0.012841	9.592	7.34e-05	u가 1 증가하면 y는 0.123 증가
v	1.989017	0.016586	119.923	2.27e-11	v가 y에 매우 강하게 양의 영향
w	-2.997816	0.005421	-552.981	2.36e-15	w는 y에 강하게 음의 영향

- 모든 변수의 p-value < 0.001 → 매우 유의미함 (**)

✅ 회귀 계수 요약

- 회귀계수의 t값은 $t = \text{회귀계수(Estimate)} / \text{표준오차(Std. Error)}$

$$t = \frac{\text{회귀계수(Estimate)}}{\text{표준오차(Std. Error)}}$$

- 이 t값은 **귀무가설(H_0 : 계수 = 0)**을 검정하기 위한 통계량이야.
- t값이 크면 클수록 해당 변수는 **통계적으로 유의미**하다는 뜻이야.

4. 모델 적합도 지표

- Residual standard error:** 0.1303 (잔차의 평균 크기 → 작음)
- Multiple R-squared:** 1
- Adjusted R-squared:** 1
→ 모델이 y의 변동을 **100% 설명함** (사실상 완벽)
- F-statistic:** 103800
 - F값이 클수록 **회귀모형이 통계적으로 유의**하다는 뜻

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n - p - 1)}$$

기호	의미
SSR	회귀 제곱합 (모델이 설명한 부분)
SSE	잔차 제곱합 (오차)
p	독립변수 개수
n	전체 관측값 수
MSR	Mean Square for Regression = SSR / p
MSE	Mean Square Error = SSE / (n - p - 1)

- F-test p-value:** 1.564e-14
→ 모델 전체가 통계적으로 매우 유의미함

💡 해석 요약

- 이 모델은 거의 “완벽” 수준으로 y를 예측함
- 변수 **v**와 **w**는 영향력이 **매우 큼**, **u**도 영향 있음
- 오차도 거의 없음 → 실험 데이터이거나 과적합일 가능성도 염두에 두어야 함
- 실제 데이터라면 거의 **설명력이 100%인 드문 경우임**

데이터셋 ChickWeight

✓ 요약

- **ChickWeight** 는 병아리의 성장 데이터를 담은 R 내장 데이터셋으로, 병아리에게 서로 다른 사료를 주며 시간에 따른 몸무게 변화를 측정한 실험 결과야.
- 반복 측정 자료(repeated measures) 형태이기 때문에 혼합효과모델, 로지스틱 회귀, 선형모형 분석 등 다양한 실습에 사용됨.
- 총 578개의 관측값, 4개 변수, 50마리의 병아리 데이터를 포함하고 있음.

ChickWeight 데이터셋 상세 설명

1. 데이터 구조

```
str(ChickWeight)
```

출력 예시:

```
'data.frame': 578 obs. of 4 variables:
 $ weight: num 42 51 59 64 76 93 106 125 149 171 ...
 $ Time : num 0 2 4 6 8 10 12 14 16 18 ...
 $ Chick : Ord.factor w/ 50 levels "1"<"2"<"3"<...
 $ Diet : Factor w/ 4 levels "1","2","3","4"
```

2. 각 변수 설명

변수명	설명	예시
weight	병아리의 몸무게 (그램)	42, 51, 59, ...
Time	실험 경과일 (일 단위)	0, 2, 4, ..., 21
Chick	병아리 개체 식별자 (총 50마리)	"1", "2", ..., "50"
Diet	사료 종류 (1~4번)	"1", "2", "3", "4"

실험 디자인 요약

- 총 50마리 병아리가 4가지 서로 다른 사료(Diet) 그룹 중 하나에 배정됨
- 각 병아리는 **여러 시점(Time)**에서 **몸무게(weight)**를 측정함
- 결과적으로 반복 측정(같은 병아리의 다회 측정) 구조를 가짐

코드

```
model1 = lm(y ~ x, df)
plot(model1)
```

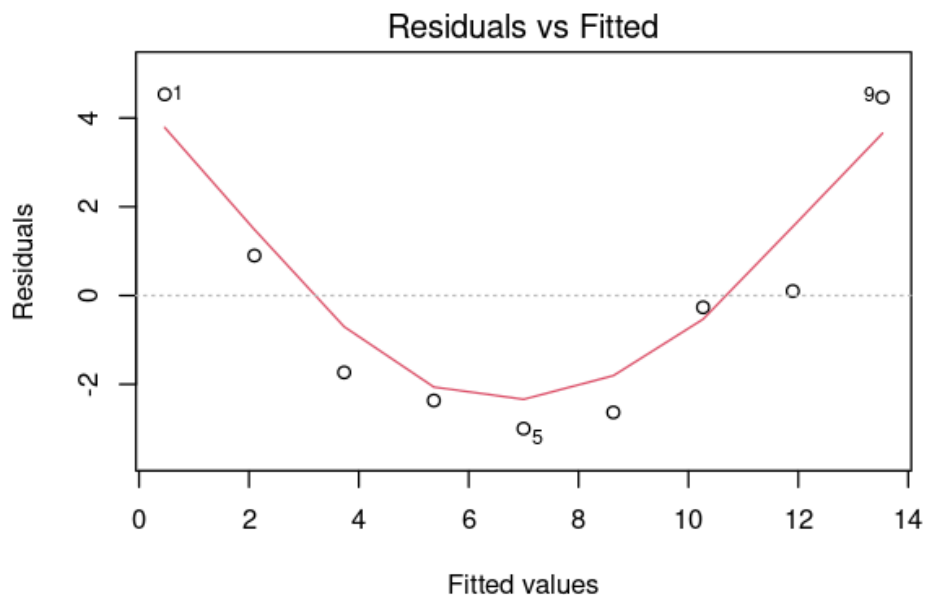
✅ 요약

- `plot(lm_object)` 은 회귀모델 진단용 기본 4개의 플롯을 출력함
- 이 그래프들은 각각 잔차, 정규성, 레버리지, 이상치 등을 시각적으로 보여줌
- 이 플롯들을 통해 선형 회귀모형이 적절한지 여부를 직관적으로 파악할 수 있음

1. plot(lm)에서 나오는 4가지 플롯 설명

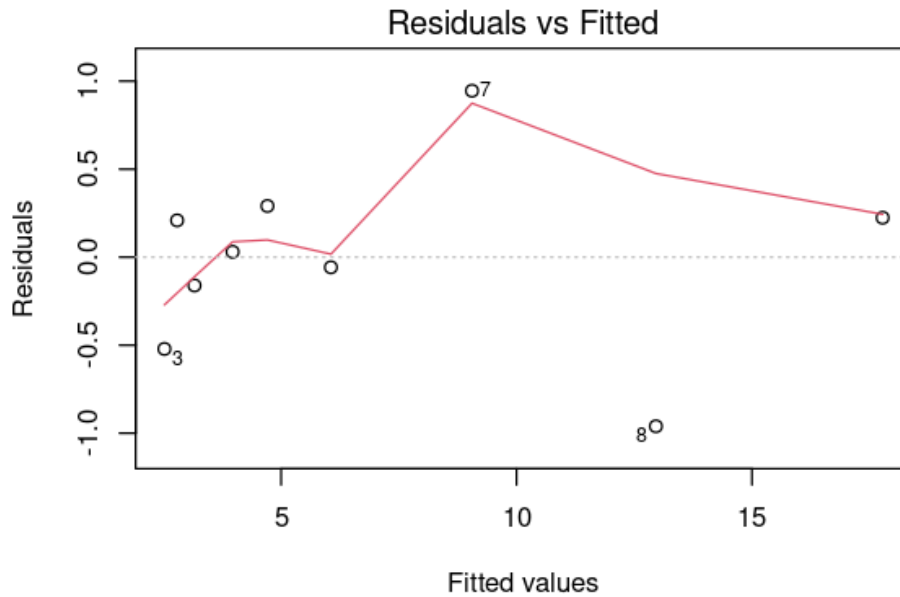
① Residuals vs Fitted (잔차 vs 예측값)

- X축: 예측값 (y^{\wedge})
- Y축: 잔차 ($y - y^{\wedge}$)
- 목적: 선형성(linearity) 및 등분산성(homoscedasticity) 점검
- 해석 포인트:
 - 랜덤하게 흩어진 패턴이면 OK
 - 곡선형태나 깔때기 형태 → 비선형/이분산성 의심됨



식을 x 뿐만 아니라 x^2 를 추가해서 모델 model2를 만들어 그려보니

- 결과가 좋음



② Normal Q-Q (잔차의 정규성 검정)

- **X축:** 이론적 분위수 (정규분포)
- **Y축:** 관측된 잔차의 분위수
- **목적:** 오차가 정규분포를 따르는지 시각화
- **해석 포인트:**
 - 점들이 **45도 직선**에 잘 붙어 있으면 OK
 - 끝부분에서 튀면 **정규성 위배**

③ Scale-Location (Spread-Location)

- **X축:** 예측값
- **Y축:**

$$\sqrt{|e|} \text{ 또는 } \sqrt{\text{표준화잔차}}$$

- **목적:** 오차의 분산이 일정한지 확인 (등분산성)
- **해석 포인트:**
 - 평평한 수평선이면 OK
 - 점들이 위나 아래로 퍼지는 패턴 → 이분산성 문제 있음

④ Residuals vs Leverage (레버리지 vs 잔차)

- **X축:** 레버리지 (influence)
- **Y축:** 잔차
- **목적:** 이상치(outlier) 및 영향점(influential point) 탐지

- 해석 포인트:

- 점이 Cook's distance 선 바깥에 있음 → 영향력 큰 이상치!
- 레버리지가 큰 관측값은 주의 깊게 봐야 함

2. 참고: plot이 한 번에 안 뜰 때는?

R 콘솔에서 한 번에 4개가 안 뜰 수도 있으니 이럴 땐:

```
par(mfrow = c(2, 2))  
plot(model1)
```

코드

```
> stepAIC(Y ~ X1+X2+X3+X4, df,  
+ scope = list(lower = ~1, upper = ~X1+X2+X3+X4), direction = "backward")  
Start: AIC=26.94  
Y ~ X1 + X2 + X3 + X4  
  
Df Sum of Sq  RSS   AIC  
- X3   1    0.1091 47.973 24.974  
- X4   1    0.2470 48.111 25.011  
- X2   1    2.9725 50.836 25.728  
<none>          47.864 26.944  
- X1   1   25.9509 73.815 30.576  
  
Step: AIC=24.97  
Y ~ X1 + X2 + X4  
  
Df Sum of Sq  RSS   AIC  
<none>          47.97 24.974  
- X4   1     9.93 57.90 25.420  
- X2   1    26.79 74.76 28.742  
- X1   1   820.91 868.88 60.629  
  
Call:  
lm(formula = Y ~ X1 + X2 + X4, data = df)  
  
Coefficients:  
(Intercept)      X1      X2      X4  
    71.6483    1.4519    0.4161   -0.2365
```



- 처음에는 ****모든 변수(X1, X2, X3, X4)****를 포함한 전체 모델로 시작했어.
- 이후 **AIC**가 가장 많이 감소하는 변수를 하나씩 제거하면서
X3가 제거되고, 최종 모델은 **X1 + X2 + X4**로 결정됐어.
- 최종 모델은 AIC = 24.97로 이전보다 더 간결하면서도 성능을 유지하는 모델이야.

1. step() 함수 요약

```
step(lm(Y ~ X1+X2+X3+X4, df),
  scope = list(lower = ~1, upper = ~X1+X2+X3+X4),
  direction = "backward")
```

- **direction = "backward"** : 후진 제거법
 - 처음에는 모든 변수 포함
 - AIC 기준으로 가장 덜 중요한 변수 하나씩 제거
- **scope** : 제거 가능한 범위 (최소 모델은 상수항 only, 최대 모델은 전체 변수)
 - **scope**는 ****모델 탐색의 최소 범위(lower)****와 ****최대 범위(upper)****를 지정하는 옵션
 - **lower = ~1**은 절편만 있는 모델(즉, 아무 변수도 없는 모델)
 - **upper = ~X1+X2+X3+X4**는 모든 변수 포함 모델
 - 따라서 이 **scope**는 "절편만 있는 모델부터 전체 변수 포함 모델까지"를 탐색 범위로 지정한 거야

예시 비교

설정	의미
lower = ~1	절편만 있는 모델부터 시작/도달 가능
lower = ~X1+X2	최소한 X1, X2는 항상 포함됨
upper = ~X1+X2+X3+X4	최대한 이 변수들까지만 포함 가능
upper = ~.	데이터프레임에 있는 모든 변수 포함 가능

2. 진행 과정 해석

✅ Step 1: 전체 모델로 시작

- 현재 **AIC = 26.94**
- 변수 중 제거하면 AIC가 더 줄어드는 항목 탐색

제거 변수	AIC 결과	비고
-X3	24.974	✓ 가장 낮음 → 제거됨
-X4	25.011	

제거 변수	AIC 결과	비고
-X2	25.728	
-X1	30.576	X → 제거하면 안 됨

→ X3가 가장 효과 없는 변수로 판단되어 제거됨

✓ Step 2: 모델이 $Y \sim X1 + X2 + X4$ 로 업데이트됨

- 현재 AIC = **24.974**
- 변수 제거 시 AIC 증가하므로 더 제거하지 않음

제거 변수	AIC 결과	비고
-X4	25.42	제거 시 AIC 증가 ❌
-X2	28.74	제거 시 AIC 증가 ❌
-X1	60.63	제거 시 AIC 폭증 ❌

→ 더 이상 제거할 변수 없음 → 최종 모델 확정

3. 최종 선택된 모델

```
lm(formula = Y ~ X1 + X2 + X4, data = df)
```

회귀계수

항목	계수(Estimate)	해석
(Intercept)	71.6483	y 절편
X1	1.4519	X1이 1 증가하면 y는 1.45 증가
X2	0.4161	X2가 1 증가 → y 약간 증가
X4	-0.2365	X4가 증가하면 y는 감소함

4. 모델 선택 기준: AIC란?

- AIC (Akaike Information Criterion): $AIC = 2k - 2\log(L)$
 $AIC = 2k - 2\log(L)$
 - k = 모델의 파라미터 수
 - L = 최대우도
- AIC는 작을수록 좋은 모델
 - 설명력 + 단순성 모두 고려한 균형 지표
 - 너무 많은 변수 쓰는 모델은 패널티 받음

