

# 확률통계 및 공통 알고리즘

## 5. 회귀분석

임채영

서울대학교

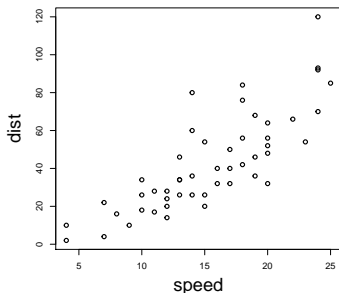
## 이번 강의에서 다룰 내용

- ▶ 단순선형회귀분석- 회귀계수의 추정, 결정계수, 회귀직선의 유의성 검정, 회귀계수의 유의성 검정, 평균반응에 대한 추론, 잔차분석
- ▶ 다중선형회귀분석 - 회귀계수의 추정, 결정계수, 유의성 검정, 범주형설명변수

# 선형 모형과 회귀분석

- ▶ 하나의 변수(설명변수, Explanatory variable)가 다른 하나의 변수(반응변수, Response variable)에 영향을 끼치는 관계를 설명하기 위한 모형을 생각해 볼 수 있다.
- ▶ 설명변수와 반응변수 간에는 다양한 함수 형태의 관계를 가정할 수 있는데, 특히 반응변수가 입력변수의 1차 함수 형태로 나타난다고 가정하는 경우, 선형 모형(Linear model)을 고려해볼 수 있다.

# 선형 모형 예시



자동차의 속도와 정지 시 제동 거리로 이루어진 자료의 산점도.

- ▶ '자동차의 제동 거리는 속도에 대한 직선 형태의 함수로 나타날 것이다' 라고 생각하면, 두 변수 간에 다음과 같은 선형 모형을 고려할 수 있다.
- ▶  $\text{거리} = \beta_0 + \text{속도} \times \beta_1$

# 회귀분석

- ▶ 회귀분석 (Regression Analysis)

변수사이의 모형을 가정하고 자료를 이용하여 추정하는 분석 방법

- ▶ 선형회귀모형 (Linear regression model)

선형모형을 가정한 회귀모형

# 단순 선형회귀모형

- ▶ 단순선형회귀모형(Simple linear regression model)

하나의 설명변수에 대해 만들어진 선형 회귀모형

$$y = \beta_0 + \beta_1 x + \epsilon, E(\epsilon) = 0, Var(\epsilon) = \sigma^2$$

- ▶ 회귀계수(Regression coefficient)

모형의 계수들 ( $\beta_0, \beta_1$ )

## 단순 선형회귀모형의 가정

- ▶ 데이터  $(Y_i, X_i)$ 가 독립적으로 관측되었을때

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

$$\epsilon_i \stackrel{i.i.d.}{\sim} (0, \sigma^2)$$

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

- ▶ 독립성, 선형성, 등분산성
- ▶ 추가로 오차의 분포가 정규분포라는 가정을 통해,  
모회귀계수에 대한 추론을 할 수 있게 된다. (정규성 가정)

# 회귀계수의 추정

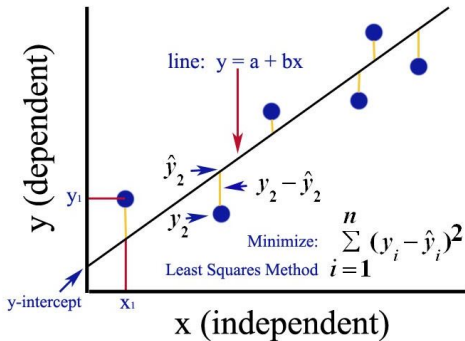
- ▶ 최소제곱법(method of least square)

오차 제곱합  $\sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$ 을 최소화시키는  $\beta_0, \beta_1$ 를 추정

- ▶ 이렇게 구해진 추정량을 최소제곱추정량(Least Square Estimator, LSE)이라 한다.



# 최소제곱법의 원리



- ▶ 구체적으로 오차의 제곱합  $\sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$  을  $\beta_0, \beta_1$ 에 대해 각각 편미분하여 얻은 식을 0으로 놓은 연립방정식 (정규 방정식, normal equation)의 해를 구함

$$\begin{cases} -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ -2 \sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases}$$

- ▶  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$

- ▶ 최소제곱회귀직선

최소제곱법으로 추정한 회귀계수를 대입하여 만든 직선

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = \bar{Y} + \hat{\beta}_1 (X - \bar{X})$$

- ▶  $\hat{Y}_i$ : 반응변수의 기댓값인  $E(Y_i|X_i)$ 의 추정값, 즉,  
 $\hat{Y}_i = \widehat{E(Y_i|X_i)}$

# 예시

20명의 수학능력시험의 국어영역과 영어영역의 점수가 아래와 같이 주어졌다고 하자.

학생 번호	1	2	3	4	5	6	7	8	9	10
국어	42	38	51	53	40	37	41	29	52	39
영어	30	25	34	35	31	29	33	23	36	30
학생 번호	11	12	13	14	15	16	17	18	19	20
국어	45	34	47	35	44	48	47	30	29	34
영어	32	29	34	30	28	29	33	24	30	30

모국어에 관련된 어학 능력이 외국어에 관련된 어학 능력에 영향을 끼치는지 알아보기 위해, 국어 점수를 설명변수로 하고 영어 점수를 반응변수로 하는 단순회귀모형을 적합해보도록 한다.

- ▶ X와 Y를 각각 국어성적과 영어성적이라 했을 때, 통계량 및 회귀계수는 다음과 같다.
- ▶  $\bar{x} = 40.75, \bar{y} = 30.25, S_{xx} = 1083.75, S_{xy} = 379.25$
- ▶  $\hat{\beta}_1 = \frac{379.25}{1083.75} = 0.3499, \hat{\beta}_0 = 30.25 - \hat{\beta}_1 \cdot 40.75 = 15.9899$
- ▶ 즉, 적합된 회귀모형은 다음과 같다.

$$\hat{Y}_i = 15.9899 + 0.3499 \times x_i$$

## ▶ Python statmodels 모듈을 사용한 회귀분석 시행 결과

```
In [207]: kor = df(data=kor) # 모듈 및 저장시에는 pandas.DataFrame
...: eng = df(data=eng)
...: score = pd.concat([kor, eng], axis=1)
...: #axis= 0 : 한 열로 이어붙이기, 1 : 서로 다른 열로 구분해서 합치기
...: score_model=sm.ols(formula='eng~kor', data=score) # ols : 모형화 할수
...: score_result=score_model.fit() # 모형화 결과(빈 괄호부분을 생략하면 결과를 볼 수 없다.)
...: print(score_result.summary()) #결과 정리 : 결과값은 이 과정을 통해 볼 수 있다.
```

### OLS Regression Results

Dep. Variable:	eng	R-squared:	0.573
Model:	OLS	Adj. R-squared:	0.549
Method:	Least Squares	F-statistic:	24.12
Date:	Tue, 19 Dec 2017	Prob (F-statistic):	0.000113
Time:	22:13:08	Log-Likelihood:	-44.376
No. Observations:	20	AIC:	92.75
Df Residuals:	18	BIC:	94.74
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	15.9899	2.950	5.419	0.000	9.791	22.189
kor	0.3499	0.071	4.911	0.000	0.200	0.500

Omnibus:	1.491	Durbin-Watson:	2.023
Prob(Omnibus):	0.474	Jarque-Bera (JB):	1.301
Skew:	-0.521	Prob(JB):	0.522
Kurtosis:	2.310	Cond. No.	233.

## 잔차 (residual)

- ▶ 잔차(residual)

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

- ▶ 오차  $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ 의 관측값 개념으로 생각할 수 있고, 이들은 오차분산  $\sigma^2$ 의 크기에 따라 값이 크거나 작게 나타남
- ▶ 잔차의 크기가 작을 수록 최소제곱회귀직선이 실제 관측 결과를 잘 설명해준다고 생각할 수 있음

## 잔차제곱합 (Residual sum of square)

- ▶ 잔차제곱합(Residual sum of square, SSE)

$$SSE(= SS_{Res}) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- ▶ 오차분산  $\sigma^2$  추정에 사용

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- $MSE$  (평균제곱오차, Mean squared error): SSE를 잔차제곱합의 자유도 (표본의 수 - 회귀계수의 수)로 나눈값

## 편차의 분해

자료를 적합시킨 선형회귀모형이 관측자료를 얼마나 잘 설명하고 있는지 판단하기 위해 고려함

- ▶ 편차,  $y_i - \bar{y}$ : 관측값  $y_i$ 와 평균  $\bar{y}$ 의 차이
  - $x_i$  (설명변수)가 없을때  $y_i$  (반응변수)를 설명하는  $\bar{y}$ 와의 차이
- ▶ 편차의 분해

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = \hat{\epsilon}_i + (\hat{y}_i - \bar{y})$$

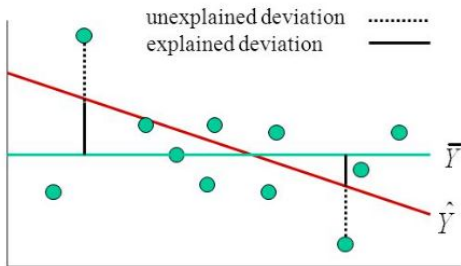
- $(y_i - \hat{y}_i)$ : 잔차, 회귀직선으로 설명하지 못하는 부분
- $(\hat{y}_i - \bar{y})$ 는 회귀직선에 의해 설명되는 부분
- ▶ 두 종류의 편차를 더했다는 뜻에서  $y_i - \bar{y}$ 를 총편차(total deviation)라고 부른다.



# 편차의 분해

그림으로 나타낸 편차의 분해

- ▶ 실선: 회귀직선에 의해 설명되는 편차
- ▶ 점선: 잔차



## 편차의 제곱합

- ▶ 반응변수 값의 변동 중 회귀직선을 통해 설명할 수 있는 변동을 구할수 있음
- ▶ 편차의 합은 그 특성상 무조건 0이 되므로, 편차의 제곱합을 사용

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

- 총제곱합 (Total sum of square, SST): 총편차의 제곱합, 자료 전체의 변동
- 회귀제곱합 (Regression sum of square, SSR): 회귀직선을 통해 설명할 수 있는 편차의 제곱합
- 잔차제곱합(SSE or  $SS_{Res}$ )

## 결정계수, $R^2$

- ▶ 결정계수(Coefficient of determination,  $R^2$ )

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

- 총 제곱합 중 회귀제곱합이 차지하는 비율
- ▶ 총제곱합 중 회귀제곱합의 부분이 클수록 회귀모형이 관측결과를 잘 설명해주는 것이라 할 수 있다.

## 단순선형회귀모형에서의 결정계수

$$R^2 = \frac{SSR}{SST} = \frac{(S_{xy})^2}{S_{xx} S_{yy}} = \left\{ \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}} \right\}^2 = r^2$$

- ▶ 단순선형회귀모형

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \{\hat{\beta}_1(x_i - \bar{x})\}^2 = \frac{(S_{xy})^2}{S_{xx}}$$

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

- ▶ 결정계수  $R^2$ 은 표본상관계수  $r$ 의 제곱
- ▶  $R^2$ 이 1에 가까울수록 산점도에서 점들이 직선 주위에 밀집되어 나타난다고 볼 수 있다. 즉, 회귀직선이 자료를 잘 설명한다고 해석할 수 있다.

## 예시

- ▶ 앞에서 다룬 국어 성적과 영어 성적의 편차 제곱합을 구해보자.
- ▶  $S_{xx} = 1083.75$ ,  $S_{xy} = 379.25$ ,  $S_{yy} = 231.75$
- ▶  $SST = S_{yy} = 231.75$ ,  $SSR = \frac{(S_{xy})^2}{S_{xx}} = 132.716$   
 $SSE = SST - SSR = 99.034$
- ▶ 총제곱합과 회귀제곱합을 통해 구한 결정계수와 앞서 구한 상관계수의 관계를 구해보자.

$$R^2 = \frac{132.716}{231.75} = 0.573 = (0.757)^2 = r^2$$

- ▶ 즉, 앞서 적합한 단순선형회귀모형은 자료의 변동의 약 57.3%를 설명한다고 볼 수 있고, 결정계수가 상관계수의 제곱으로 나타난다는 것 역시 확인해보았다.

## 단순회귀분석에서의 추론: 회귀직선의 유의성 검정

- ▶ 설명변수  $x$ 를 고려한 회귀직선이 의미가 있는지에 대한 검정
- ▶  $\frac{SSR}{SSE}$ 의 값이 커질 수록 회귀직선의 유의하다는 증거가 강해지는것으로 해석
- ▶  $SSR/\sigma^2 \sim \chi^2(1)$ ,  $SSE/\sigma^2 \sim \chi^2(n-2)$
- ▶  $F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F(1, n-2)$

# F검정

- ▶ 귀무가설:  $H_0$  : 회귀모형이 유의하지 않다. ( $\beta_1 = 0$ )
- ▶ 대립가설  $H_1$  : 회귀모형이 유의하다. ( $\beta_1 \neq 0$ )
- ▶ 검정통계량 :  $F = \frac{SSR/1}{SSE/(n-2)}$
- ▶ 검정통계량의 관측값  $f_0$ .
- ▶ 유의확률:  $P(F \geq f_0)$
- ▶ 유의수준  $\alpha$ 에서 기각역 :  $F > F_{1-\alpha}(1, n-2)$
- ▶  $F \sim F(k_1, k_2)$ 일때,  $P(F \leq F_p(k_1, k_2)) = p$ .

# 분산분석표

## ▶ 분산분석표(Analysis of variance table)

회귀직선의 유의성 검정 과정을 표로 요약

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	$SSR$	1	$MSR = SSR / 1$	$f_0 = MSR / MSE$	$P(F \geq f_0)$
잔차	$SSE$	$n - 2$	$MSE = SSE / (n - 2)$		
계	$SST$	$n - 1$			



# 예시

앞서 국어성적을  $x$ 로 설정하고 영어성적을  $y$ 로 설정하여 적합한 회귀직선  $Y_i = 15.9899 + 0.3499 \times x_i + \epsilon_i$ 에 대한 F검정을 해보자.

▶ 회귀모형에 대한 가설

$H_0$  : 회귀모형이 유의하지 않다. ( $\beta_1 = 0$ )

$H_1$  : 회귀모형이 유의하다. ( $\beta_1 \neq 0$ )

	제곱합	자유도	평균제곱	F값	유의확률
회귀	132.716	1	132.716	24.122	<0.001
잔차	99.034	18	5.502		
계	231.75				

▶ 분산분석표를 관측한 결과, 귀무가설을 기각할 수 있음을 확인할 수 있다. 즉, 적합한 회귀직선은 유의하다고 할 수 있다.

- ▶ statsmodels 패키지 안의 stats 서브패키지에 있는 anova 모듈 안의 *anova\_lm* 함수를 활용하여 분산분석표를 얻을 수 있다.

```
In [208]: anova.anova_lm(score_result)
C:\Users\user\Anaconda3\lib\site-packages\scipy\stats\_distn_infrastructure.py:879: RuntimeWarning: invalid
value encountered in greater
    return (self.a < x) & (x < self.b)
C:\Users\user\Anaconda3\lib\site-packages\scipy\stats\_distn_infrastructure.py:879: RuntimeWarning: invalid
value encountered in less
    return (self.a < x) & (x < self.b)
C:\Users\user\Anaconda3\lib\site-packages\scipy\stats\_distn_infrastructure.py:1821: RuntimeWarning: invalid
value encountered in less_equal
    cond2 = cond0 & (x <= self.a)
Out[208]:
```

	df	sum_sq	mean_sq	F	PR(>F)
kor	1.0	132.715629	132.715629	24.12174	0.000113
Residual	18.0	99.034371	5.501910	NaN	NaN

## 회귀계수 ( $\beta_1$ )의 유의성에 대한 검정

- ▶ 회귀계수의 최소제곱추정량의 성질을 이용하여 회귀계수의 유의성에 대한 검정 가능
- ▶ 단순선형회귀에서

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} / \sqrt{S_{xx}}} \sim t(n-2)$$

# 회귀계수 유의성검정을 위한 t 검정

- ▶ 귀무가설:  $H_0$  : 회귀계수가 유의하지 않다. ( $\beta_1 = b$ )
- ▶ 대립가설  $H_1$  : 회귀계수가 유의하다. ( $\beta_1 \neq b$ )
- ▶ 검정통계량 :  $t = \frac{\hat{\beta}_1 - b}{\hat{\sigma} / \sqrt{S_{xx}}}$
- ▶ 검정통계량의 관측값:  $t_0$

$T \sim t(k)$ 일때,  $P(T \leq t_p(k)) = p$ .

대립가설 $H_1$	유의확률	유의수준 $\alpha$ 의 기각역
$\beta_1 > b$	$P(T > t_0)$	$T > t_{1-\alpha}(n-2)$
$\beta_1 < b$	$P(T < t_0)$	$T < -t_{1-\alpha}(n-2)$
$\beta_1 \neq b$	$P( T  >  t_0 )$	$ T  > t_{1-\alpha/2}(n-2)$

## 예시

중고차의 사용 연수에 따라 중고차의 가격이 얼마나 달라지는 지에 대해 알아보기 위해, 중고차의 사용 연수를 설명변수( $x$ , 단위 : 1년), 가격을 반응변수( $y$ , 단위 : 100만원)로 하는 단순선형회귀모형을 고려해보자

사용연수 ( $x_i$ )	1.0	1.5	2.0	2.0	3.0	3.0	3.2	4.0	4.5	5.0	5.0	5.5
중고차가격 ( $y_i$ )	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4

중고차의 사용 연수 및 가격 자료

- ▶  $n = 12$ ,  $\bar{x} = 3.308$ ,  $\bar{y} = 2.292$   
 $S_{xx} = 24.649$ ,  $S_{xy} = -21.169$ ,  $\hat{\sigma}^2 = MSE = 0.0289$
- ▶  $\hat{\beta}_1 = \frac{-21.169}{24.649} = -0.859$ ,  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5.133$
- ▶ 적합된 회귀모형:

$$\hat{Y}_i = 5.133 - 0.859 \times x_i$$

적합된 회귀모형의 회귀계수  $\beta_1$  이 유의한지에 대해 유의수준 5%에서  
t 검정을 시행해보자.

- ▶ 회귀계수에 대한 가설

$H_0$  : 회귀계수  $\beta_1$  이 유의하지 않다. ( $\beta_1 = 0$ )

$H_1$  : 회귀계수  $\beta_1$  이 유의하다. ( $\beta_1 \neq 0$ )

- ▶ 검정통계량의 관측값 :

$$t_0 = \frac{\hat{\beta}_1 - 0}{\hat{\sigma} / \sqrt{S_{xx}}} = \frac{-0.859}{\sqrt{0.0289/24.649}} = -25.087$$

- ▶ 기각역:  $|t_0| > |t_{0.975}(10)| = 2.281$

- ▶ 유의수준 5%에서 회귀계수가 유의하다고 볼 수 있다.

## summary 명령어를 통해 나온 결과로부터 계수의 유의성검정 진행

```
In [209]: usedcar_model=sm.ols(formula='price~year', data=usedcar) # ols : 모형화 함수  
...: usedcar_result=usedcar_model.fit() # 모형화 결과(반 괄호부분을 생략하면 결과를 볼 수 없다.)  
...: print(usedcar_result.summary()) #결과 정리 : 결과값은 이 과정을 통해 볼 수 있다.
```

### OLS Regression Results

```
=====
Dep. Variable:          price    R-squared:                0.984
Model:                  OLS      Adj. R-squared:            0.983
Method:                 Least Squares    F-statistic:        629.8
Date:                  Tue, 19 Dec 2017  Prob (F-statistic):    2.31e-10
Time:                  22:16:16    Log-Likelihood:        5.3366
No. Observations:      12         AIC:                   -6.673
Df Residuals:          10         BIC:                   -5.703
Df Model:               1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.1329	0.123	41.600	0.000	4.858	5.408
year	-0.8588	0.034	-25.095	0.000	-0.935	-0.783

```
=====
Omnibus:                 0.703    Durbin-Watson:           1.413
Prob(Omnibus):           0.704    Jarque-Bera (JB):        0.588
Skew:                    0.073    Prob(JB):                0.745
Kurtosis:                 1.926    Cond. No.                 9.66
=====
```

## 예시

이번에는 회귀계수의 값, 즉 연 평균 중고차 가격의 감소액이 80 만원을 초과하는지에 대해 검정해보도록 한다.

- ▶ 회귀계수에 대한 가설

$$H_0 : \beta_1 = -0.8, H_1 : \beta_1 < -0.8$$

- ▶ 검정통계량의 관측값

$$t_0 = \frac{\hat{\beta}_1 - (-0.8)}{\hat{\sigma} / \sqrt{S_{xx}}} = \frac{-0.859 - (-0.8)}{\sqrt{0.0289/24.649}} = -1.719$$

- ▶ 기각역 :  $t_0 < -t_{0.95}(10) = -1.812$
- ▶ 통계량의 관측값이 기각역에 들어가지 못하므로, 유의수준 5%에서 귀무가설을 기각하지 못한다.



## 예시: $\beta_1$ 의 신뢰구간

- ▶ 앞서 구한  $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ 를 이용하여 회귀계수  $\hat{\beta}_1$ 에 대한 95% 신뢰구간을 구할 수 있다

$$\hat{\beta}_1 \pm t_{0.975}(12-2) \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = -0.859 \pm 2.228 \cdot \sqrt{\frac{0.0289}{24.649}}$$

$$= -0.859 \pm 0.076 \Rightarrow (-0.935, -0.783)$$

- ▶ 중고차의 가격은 1년이 지날 때마다 평균적으로 783,000원에서 935,000 사이로 감소한다고 신뢰수준 95%에서 결론내릴 수 있다.

## t검정과 F검정의 관계

- ▶ 단순선형회귀분석에서,  
회귀모형의 유의성에 대한  $F$  검정 = 회귀계수의 유의성에 대한  $t$  검정

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{\hat{\beta}_1^2 S_{xx}}{MSE} = \left\{ \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{S_{xx}}} \right\}^2 = T^2$$

- ▶  $T \sim t(n-2) \rightarrow T^2 \sim F(1, n-2)$

## 회귀직선의 상수항 ( $\beta_0$ ) 에 대한 추론

- ▶ 상수항이 어떤값을 갖는지에 대한 유의성 검정을 최소제곱법 추정량  $\hat{\beta}_0$ 의 성질을 이용하여 할수 있다.
- ▶  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ 로부터
- ▶  $E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1 \bar{x}) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$
- ▶  $Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x}) =$   
 $Var(\bar{y}) + (\bar{x})^2 var(\hat{\beta}_1) - 2Cov(\bar{y}, \hat{\beta}_1 \bar{x})$   
 $= \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left( \frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}} \right)$   
( $\because Cov(\bar{y}, \hat{\beta}_1 \bar{x}) = \bar{x} Cov(\bar{y}, \hat{\beta}_1) = 0$ 임이 알려져 있다.)
- ▶  $\hat{\beta}_0 \sim N \left( \beta_0, \sigma^2 \left( \frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}} \right) \right)$
- ▶  $T = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}} \sim t(n-2)$

▶ 귀무가설  $H_0 : \beta_0 = b$

▶ 검정 통계량:  $T = \frac{\hat{\beta}_0 - b}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}}$

▶ 검정통계량의 관측값:  $t_0$

$T \sim t(k)$ 일때,  $P(T \leq t_p(k)) = p$ .

대립가설 $H_1$	유의확률	유의수준 $\alpha$ 의 기각역
$\beta_0 > b$	$P(T > t_0)$	$T > t_{1-\alpha}(n-2)$
$\beta_0 < b$	$P(T < t_0)$	$T < -t_{1-\alpha}(n-2)$
$\beta_0 \neq b$	$P( T  >  t_0 )$	$ T  > t_{1-\alpha/2}(n-2)$

$\beta_0$ 의  $100(1 - \alpha)\%$  신뢰구간

$$(\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{n^{-1} + \bar{x}^2/S_{xx}}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{n^{-1} + \bar{x}^2/S_{xx}})$$

## 잔차분석 (Residual analysis)

- ▶ 잔차의 분포를 분석하여 단순선형회귀모형에 대한 가정이 타당한지 점검하여 모형의 적정성 여부를 검토
- ▶ 단순회귀모형의 가정

$$\begin{cases} Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ (선형성)} \\ \epsilon_i \sim \mathbf{N}(0, \sigma^2) \text{이고, 서로 독립 (독립성, 등분산성, 정규성)} \end{cases}$$

# 스튜던트화 잔차

- ▶ 오차항의 성질

$$\frac{\epsilon_i}{sd(\epsilon_i)} = \{Y_i - (\beta_0 + \beta_1 x_i)\} / \sigma \sim_{iid} N(0, 1), \quad i = 1, 2, \dots, n$$

- ▶ 스튜던트화 잔차(Studentized residual):

$$\hat{e}_{st,i} = \hat{e}_i / \hat{sd}(\hat{e}_i)$$

- ▶ 스튜던트화 잔차를 마치 표준화된 오차항의 관측값처럼 생각하여 회귀모형의 가정을 검토.

# 잔차도

- ▶ 잔차도 (Residual plot)

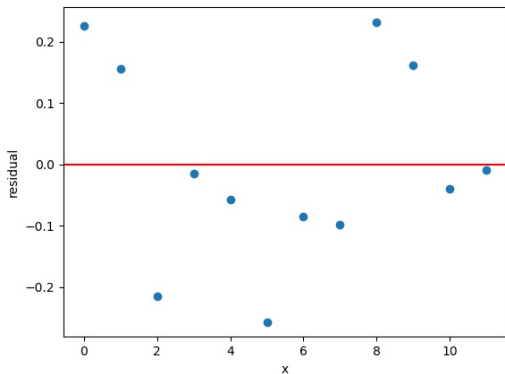
설명변수의 값들을 x좌표로, 스튜던트화 잔차를 y좌표로 그린 산점도

- ▶ 잔차도를 통해 잔차의 분포를 파악하여 회귀모형의 가정이 맞는지 판단.

- 대략 0에 관하여 대칭적으로 나타난다.
- 설명변수의 값에 따라 잔차의 산포가 크게 달라지지 않는다.
- 점들이 특정한 형식을 가지고 나타나지 않는다.
- 거의 모든 점이  $\pm 2$  범위 내에 나타난다.

# 예시

- ▶ 중고차의 사용 연수와 가격에 대한 자료를 이용하여 잔차도를 그린 결과





# 다중회귀분석 (Multiple regression analysis)

- ▶ 두 개 이상의 설명변수의 영향을 받는 경우
- ▶ 두 개 이상의 설명변수를 사용하는 선형회귀모형 :  
다중회귀선형모형(Multiple linear regression model)
- ▶  $k$ 개의 설명변수를 갖는 선형 모형의 가정

$$\begin{cases} Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ \epsilon_i \sim iid \mathbf{N}(0, \sigma^2) \end{cases}$$

- ▶  $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 는 회귀계수,  $\sigma^2$  오차분산

- ▶ 최소제곱법을 이용하여  $\beta_j$  추정

$$\sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})\}^2$$

- ▶ 위의 식을 최소화하는 회귀계수들을 편미분을 이용하여 만들어진 정규방정식을 이용하여 구함

## 회귀계수 추정량

- ▶ 선형회귀모형의 행렬을 이용한 식:  $Y = X\beta + \epsilon$

$$Y = (y_1, y_2, \dots, y_n)^T$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$$

$$\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$$

- ▶ 최소제곱법을 통한 회귀계수의 최소제곱추정량

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

# 다중회귀분석에서의 회귀직선 유의성 검정

- ▶ 귀무가설:
- ▶  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
- ▶ 대립가설:  $H_1$  : 적어도 하나의 회귀계수는 0이 아니다.
- ▶ 단순회귀분석에서의 회귀직선의 유의성 검정과 마찬가지로  $F = MSR/MSE$ 를 이용

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	$SSR$	$k$	$MSR = SSR / k$	$MSR / MSE$	$P(F \geq f)$
잔차	$SSE$	$n - k - 1$	$MSE = SSE / (n - k - 1)$		
계	$SST$	$n - 1$			

## 다중회귀분석의 결정계수

전체 자료의 변동중에 회귀직선으로 설명되는 변동의 비율

- ▶  $R^2 = \frac{SSR}{SST}$
- ▶ 결정계수는 설명변수의 개수  $k$ 에 대한 증가함수로, 설명변수의 유의성 여부와 관계없이 설명변수의 개수가 많아지면 그 값이 증가하게 되는 단점
- ▶ 단점을 보완하기 위해 수정된 결정계수(Adjusted R-Square,  $R_{adj}$ )를 사용

$$R_{adj} = 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

## 예시 - 붓꽃 데이터 (iris dataset)

- ▶ 150 송이의 붓꽃들의 꽃받침(Sepal) 길이, 꽃받침 너비, 꽃잎(Petal) 길이, 꽃잎 너비, 품종을 기록한 dataset
- ▶ 붓꽃의 다른 수치형 변수들(꽃잎 너비, 꽃받침 길이, 꽃잎 길이)이 붓꽃의 꽃받침 너비에 영향을 주는지 알아보기 위해, 다중회귀선형모형을 적합해보도록 하자.

## 회귀모형을 적합 결과

```
In [407]: iris_result = sm.ols(formula='SepalWidth ~ SepalLength+PetalLength+PetalWidth', data=iris).fit()  
...: print(iris_result.summary())
```

```
OLS Regression Results  
=====
```

Dep. Variable:	SepalWidth	R-squared:	0.524
Model:	OLS	Adj. R-squared:	0.514
Method:	Least Squares	F-statistic:	53.58
Date:	Wed, 20 Dec 2017	Prob (F-statistic):	2.06e-23
Time:	16:35:24	Log-Likelihood:	-32.100
No. Observations:	150	AIC:	72.20
Df Residuals:	146	BIC:	84.24
Df Model:	3		
Covariance Type:	nonrobust		

```
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.0431	0.271	3.855	0.000	0.508	1.578
SepalLength	0.6071	0.062	9.765	0.000	0.484	0.730
PetalLength	-0.5860	0.062	-9.431	0.000	-0.709	-0.463
PetalWidth	0.5580	0.123	4.553	0.000	0.316	0.800

```
=====
```

Omnibus:	0.738	Durbin-Watson:	1.889
Prob(Omnibus):	0.691	Jarque-Bera (JB):	0.426
Skew:	-0.102	Prob(JB):	0.808
Kurtosis:	3.163	Cond. No.	82.1

```
=====
```

▶ 데이터로 적합한 (fitted) 회귀직선

$$\begin{aligned} \text{SepalWidth} = & 1.0431 + 0.6071 \cdot \text{SepalLength} \\ & - 0.586 \cdot \text{PetalLength} + 0.558 \cdot \text{PetalWidth} \end{aligned}$$

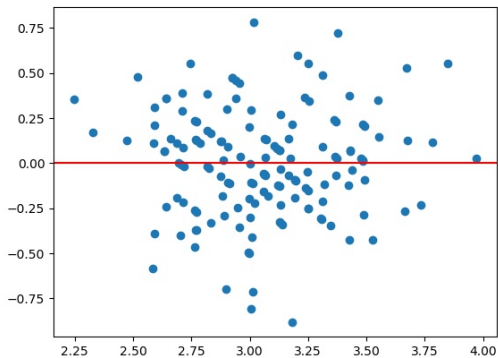
- 해석예시: 다른 설명변수가 고정되어 있을때, 꽃잎의 길이 (Petal Length)가 1 (cm)증가할때, 반응변수인 꽃받침의 너비 (Sepal Width)가 -0.586cm만큼 감소
- ▶ 회귀직선의 유의성에 대한 F 검정
  - F statistics = 53.58
- ▶ 각각의 회귀계수 및 상수항에 대한 t검정
  - 모두 유의하게 나타남
- ▶ 결정계수,  $R^2=0.524$



# 잔차분석

- ▶ 설명변수의 개수가 여러개이므로, 설명변수를  $x$ 좌표로 두고 잔차를 나타내는 것이 적절하지 않다.
- ▶ 대신 평균반응의 추정값  $\hat{y}$ 를  $x$ 좌표로 사용

## 잔차도 예시



## 범주형 설명변수

- ▶ 설명변수가 범주형 변수(categorical variable)인 경우
- ▶ Python이나 R과 같은 통계분석 프로그램에서 회귀모형을 적합할때 내부적으로 지시변수(Indicator variable) 개념 사용
- ▶ 지시변수란, 범주에 따라 0과 1 중 하나의 값을 배정하는 정수 형태의 변수
- ▶ 범주의 갯수가  $p$ 개일 때,  $p - 1$ 개의 지시변수 필요

## 범주형 설명변수 예시

- ▶ 붓꽃자료에서 범주형 변수 품종 (Species)의 경우 3개의 범주 (Setosa, Versicolor, Virginica)가 있다.

- ▶ 품종을 추가한 선형 회귀모형

$$SepalWidth = \beta_0 + \beta_1 SepalLength + \beta_2 PetalLength + \beta_4 PetalWidth + \lambda_1 Ind_{versicolor} + \lambda_2 Ind_{virginica}$$

- ▶ 지시변수의 계수를 통해, 품종에 따라 상수항이 서로 다른 회귀직선이 된다

## 붓꽃 자료에 범주형 설명변수인 '품종'을 추가한 회귀모형을 적합해본 결과

```
In [441]: result_with_cate = sm.ols(formula='SepalWidth ~ SepalLength+PetalLength+PetalWidth+Species',  
data=iris).fit()  
...: print(result_with_cate.summary())
```

```
OLS Regression Results
```

Dep. Variable:	SepalWidth	R-squared:	0.635
Model:	OLS	Adj. R-squared:	0.622
Method:	Least Squares	F-statistic:	50.14
Date:	Wed, 20 Dec 2017	Prob (F-statistic):	7.15e-30
Time:	17:11:13	Log-Likelihood:	-12.154
No. Observations:	150	AIC:	36.31
Df Residuals:	144	BIC:	54.37
Df Model:	5		
Covariance Type:	nonrobust		

```
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.6572	0.256	6.475	0.000	1.151	2.163
Species[T.versicolor]	-1.1603	0.193	-6.003	0.000	-1.542	-0.778
Species[T.virginica]	-1.3983	0.277	-5.045	0.000	-1.946	-0.850
SepalLength	0.3778	0.066	5.761	0.000	0.248	0.507
PetalLength	-0.1876	0.083	-2.246	0.026	-0.353	-0.023
PetalWidth	0.6257	0.123	5.072	0.000	0.382	0.870

```
=====
```

Omnibus:	8.540	Durbin-Watson:	2.024
Prob(Omnibus):	0.014	Jarque-Bera (JB):	10.898
Skew:	-0.356	Prob(JB):	0.00430
Kurtosis:	4.112	Cond. No.	122.

```
=====
```

▶ 적합한 회귀직선

$$\begin{aligned} \text{SepalWidth} = & 1.6572 + 0.3778\text{SepalLength} - 0.1876\text{PetalLength} + \\ & 0.6257\text{PetalWidth} - 1.1603\text{Ind}_{\text{versicolor}} - 1.3983\text{Ind}_{\text{virginica}} \end{aligned}$$

▶ 품종이 Versicolor일때의 회귀직선

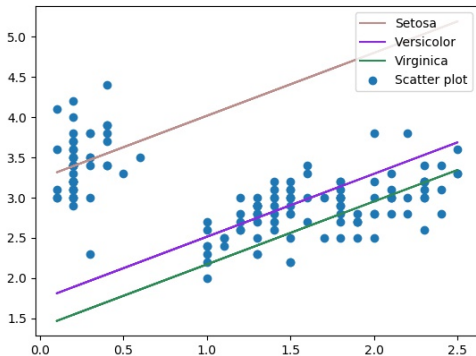
$$\begin{aligned} \text{SepalWidth} = & (1.6572 - 1.1603) + 0.3778\text{SepalLength} - \\ & 0.1876\text{PetalLength} + 0.6257\text{PetalWidth} \end{aligned}$$

▶ 품종이 Virginica일때의 회귀직선

$$\begin{aligned} \text{SepalWidth} = & (1.6572 - 1.3983) + 0.3778\text{SepalLength} - \\ & 0.1876\text{PetalLength} + 0.6257\text{PetalWidth} \end{aligned}$$

▶ 품종이 Setosa일때의 회귀직선

$$\begin{aligned} \text{SepalWidth} = & 1.6572 + 0.3778\text{SepalLength} - \\ & 0.1876\text{PetalLength} + 0.6257\text{PetalWidth} \end{aligned}$$



2차원 도표를 그리기 위해 반응변수를 '꽃받침의 너비', 설명변수를 '꽃잎의 너비'와 '품종'으로 축소시킨 회귀모형의 회귀직선. 품종에 따라 상수항이 다른 것을 확인할 수 있다.