

# 회귀분석/분산분석

서울대학교 통계연구소

2024년 2월

# 이번 강의에서 다룰 내용

- ▶ 상관계수를 이용한 상관분석
- ▶ 단순 선형회귀 분석, 예측, 모형평가
- ▶ 분산분석-일원배치

# 상관분석이란?

- ▶ 두 변수 사이의 관계 중, '직선 형태의 상관관계'에 대해 분석하는 방법을 상관분석(Correlation analysis)이라 한다.
- ▶ 두 변수 사이의 직선 관계는 산점도를 통해 대략적으로 확인할 수 있지만, 판단이 애매한 경우 보다 정량적인 척도로 상관계수(Correlation coefficient)를 사용한다.
- ▶ 상관분석에서는 상관계수의 추정량인 '표본상관계수'를 이용한다.

# 상관계수

- ▶ 상관계수  $\rho$ 의 정의는 다음과 같다.

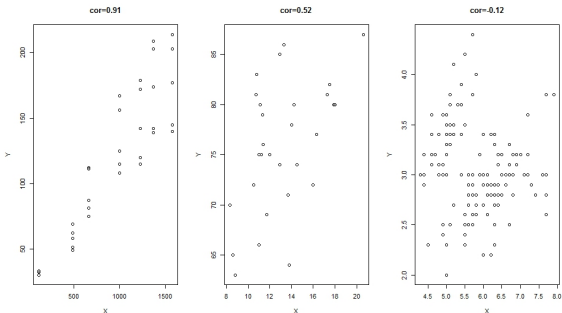
$$\rho = \text{Corr}(X, Y) = E \left[ \left( \frac{X - \mu_1}{\sigma_1} \right) \left( \frac{Y - \mu_2}{\sigma_2} \right) \right] = \frac{\text{Cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}$$

Z 정규분포라면 XY 곱의 기댓값으로 나타남

- ▶ 상관계수는 두 변수의 직선관계가 얼마나 강하고 또 어떤 방향인지를 나타내는 척도이다.
- ▶ 상관계수  $\rho$ 는 -1과 1 사이의 값을 가지며, 일반적으로 상관계수의 절댓값이 1에 가까울 수록 직선관계가 강하다고 본다. 또한 상관계수의 부호가 양이면 두 변수가 증가관계에 있다고 보고, 음이면 감소관계에 있다고 본다.

# 상관계수

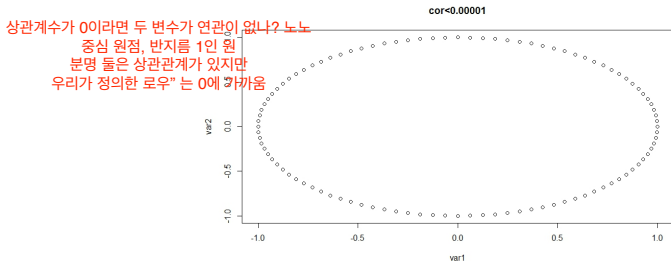
왼쪽으로 갈 수록 상관관계 커짐 (XY의 관계가 직선에 가까워짐)



상관계수의 크기에 따른 산점도의 형태, 절대값이 클 수록 선형관계가 선명하게 드러난다.

## 상관계수가 0인 경우

- ▶ 상관계수가 측정하는 두 변수 간의 상관관계는 직선 형태의 관계에 한정된다.
- ▶ 즉, 상관계수가 0이나 0에 가깝게 측정되었다는 것은, 두 변수 간의 직선관계가 드러나지 않았다는 것으로 이해해야 한다.
- ▶ 실제 두 변수 간의 상관관계가 없는 경우 상관계수가 0이 나오는 것이 사실이나, 직선이 아닌 특수한 관계를 가진 변수들끼리도 상관계수가 0이 나올 수 있다.



두 변수가 원점을 중심으로 하는 원의 x좌표와 y좌표 형태인 경우 상관계수는 0에 가까우나 변수 간의 관계가 없다고 볼 수는 없다.

# 표본상관계수

텍스트

- ▶ 표본상관계수  $\hat{\rho} = r$ 은 다음과 같이 구할 수 있다.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \left( = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \right)$$
$$= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{((\sum_{i=1}^n x_i^2 - n(\bar{x})^2)(\sum_{i=1}^n y_i^2 - n(\bar{y})^2))}}$$

- ▶ 표본상관계수 역시 -1과 1 사이의 값을 가지며,  
표본상관계수의 절댓값이 1에 가까울 수록 산점도가 직선에  
가까운 형태로 나타난다.

# 상관계수의 검정

- ▶ 모집단에서 각 개체의 두 가지 특성을 변수  $X, Y$ 로 나타낼 때, 두 변수 간의 상관계수는 모상관계수  $\rho$ 로 나타난다.
- ▶ 그리고  $(X_1, X_2, \dots, X_n)$ 과  $(Y_1, Y_2, \dots, Y_n)$ 이 각각 정규모집단으로부터의 랜덤표본일 때,  $H_0 : \rho = 0$ 에 대한 검정통계량은 다음과 같이 얻을 수 있다.  

$$T = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \sim t(n-2) \text{ under } H_0 \text{ (r: 표본상관계수)}$$
- ▶ 검정통계량의 관측값을  $t_0$ 이라 할 때, 대립가설에 따른 기각역 및 유의확률은 아래와 같이 나타난다.

$X, Y$ 는 정규분포를 따라야 함  
귀무가설 : 계수가 0 이다

대립가설	유의확률	유의수준 $\alpha$ 의 기각역
$H_1 : \rho > 0$	$P = P(T \geq t_0)$	$T \geq t_{\alpha}(n-2)$
$H_1 : \rho < 0$	$P = P(T \leq t_0)$	$T \leq -t_{\alpha}(n-2)$
$H_1 : \rho \neq 0$	$P = P( T  \geq  t_0 )$	$ T  \geq t_{\alpha/2}(n-2)$

대립가설에 종류에 따른 유의확률과 기각역의 형태



# 상관계수의 검정

예 : 다음 자료는 어느 고등학교 학생 중에서 랜덤하게 추출된 20 명의 수학능력 모의시험에서 국어영역과 영어영역의 점수이다.

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

국어영역과 영어영역 성적 간의 표본상관계수를 구해보고, 두 성적이 이변량 정규분포를 따른다고 할 때, 두 성적 사이에 상관관계가 있는지를 유의수준  $\alpha = 0.05$ 에서 검정해보자.

텍스트

# 상관계수의 검정

## R Code

```
kor <- c(42, 38, 51, 53, 40, 37, 41, 29, 52, 39, 45, 34, 47, 35, 44, 48, 47, 30, 29, 34)
eng <- c(30, 25, 34, 35, 31, 29, 33, 23, 36, 30, 32, 29, 34, 30, 28, 29, 33, 24, 30, 30)
```

```
cor(kor, eng) # 표본상관계수
```

```
[1] 0.7567
```

```
cor.test(kor, eng) # 상관분석
```

```
Pearson's product-moment correlation
```

```
data: kor and eng
```

```
t = 4.9, df = 18, p-value = 1e-04
```

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

```
95 percent confidence interval:
```

```
0.4724 0.8984
```

계산한 로의 95% 신뢰구간. [수, 수]

```
sample estimates: 0을 포함하고 있지 않기 때문에.. 또한 기각
```

```
cor
```

```
0.7567
```

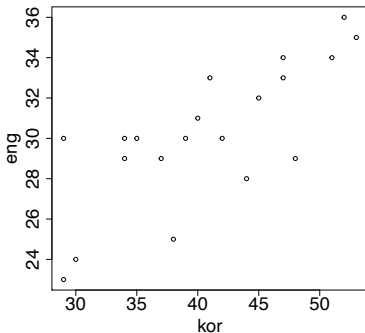
p 밸류가 매우매우 작다.  
즉 귀무가설 (로우 = 0) 을 기각  
즉 국어와 영어 성적은 상관관계가 있다.

# 상관계수의 검정

두 성적의 점수로 산점도를 그려본 결과는 다음과 같다.

```
par(cex.lab=2, cex.axis=2)  
plot(kor, eng)
```

나름 직선의 상관관계를 가지고 있다.

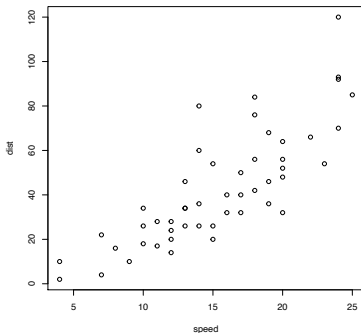


# Car 데이터

1920년대에 측정된, 자동차의 속도와 제동거리에 대한 자료

```
data(cars)  
plot(cars)
```

직선의 관계가 매우 강하다...



```
cor(cars$dist, cars$speed)
```

```
[1] 0.8069
```

```
cor.test(cars$dist, cars$speed)
```

Pearson's product-moment correlation

data: cars\$dist and cars\$speed

t = 9.5, df = 48, p-value = 1e-12

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6816 0.8862

sample estimates:

cor

0.8069

표본상관계수 = 0.8

p-값 매우 작음

귀무가설 기각

# 선형 모형과 회귀분석

여전히 두 변수가 존재,, 설명(독립)변수 + 반응(종속)변수  
일종의 함수처럼 작동  $X \rightarrow Y$  (1차함수일 경우 선형 모형)

- ▶ 앞서 살펴본 ‘상관관계’와는 다르게, 하나의 변수가 나머지 하나의 변수에 영향을 끼치는 형태의 관계들 역시 존재한다.
- ▶ 이 때, 영향을 끼치는 변수를 설명변수(Explanatory variable)라 하고, 영향을 받는 변수를 반응변수(Response variable)라 한다.
- ▶ 설명변수와 반응변수 간에는 다양한 함수 형태의 관계를 가정할 수 있는데, 특히 반응변수가 입력변수의 1차 함수 형태로 나타난다고 가정하는 경우, 이러한 관계를 설명하기 위해 선형 모형(Linear model)을 고려해볼 수 있다.

# 선형 모형

- ▶ 데이터 내의 측정된 변수들 사이의 관계를 찾는 것은 데이터 과학의 큰 주제
  - 집값은 소득의 영향을 받는가? 소득 -> 집값
  - 비료가 작물의 성장 속도를 향상시키는가? 비료 -> 성장속도
  - 키가 큰 달리기 선수가 더 빠른가? 키 -> 속도
- ▶ 이들 질문은 'x'가 'y'에 영향을 끼치는가?의 문제로 치환할 수 있다.
- ▶ 가장 간단하면서도 활용도가 높은 "영향을 끼치는 방식"으로 선형 모형을 생각할 수 있다. 즉, 'y'가 'x'의 1차 함수라고 가정하는 방법이다.

Car 데이터에서,

- ▶ 산점도를 보고, ‘자동차의 제동 거리는 속도에 대한 직선 형태의 함수로 나타날 것이다’ 와 같은 믿음을 갖게 되는 경우, 두 변수 간에 다음과 같은 선형 모델을 고려할 수 있다.
- ▶  $\text{거리} = \beta_0 + \text{속도} \times \beta_1$     속도가 거리에 영향을 끼쳤나? 얼마나 끼쳤나?
- ▶ 이 때, 선형 모델에 사용되는 계수들을 자료를 이용하여 추정하는 분석 방법을 선형 회귀분석(Linear regression analysis)이라 한다. 그리고 회귀분석을 통해 만들어진 모델을 선형회귀모형(Linear regression model)이라고도 부른다.



# 단순선형회귀

- ▶ 하나의 설명변수에 대해 만들어진 회귀모형을 단순선형회귀모형(Simple linear regression model)이라 한다.  
설명변수가 많으면 다중 (멀티플)
- ▶ 설명변수의 값  $x$ 에 대항하는 반응변수  $y$ 의 값이 직선  $\beta_0 + \beta_1 x$  주위에 나타나고, 직선에서 벗어난 측정치는 측정 오차라고 생각하는 경우, 다음과 같은 선형 모형을 생각해볼 수 있다.

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

- ▶ 이 때, 모형의 계수들을 회귀계수(Regression coefficient)라 한다.

회귀 분석의 목적

1) 예측

2) 검정 (베타1 = 0이다, 귀무가설)

$y = \text{베타}0 + \text{잔차}$ , 즉  $y$ 는  $x$ 와 관계가 없다.

$$\begin{aligned} E(y) &= E(\text{베타}0 + \text{베타}1 \cdot x + \text{잔차}) \\ &= \text{베타}0 + \text{베타}1 \cdot x + E(\text{잔차}) \end{aligned}$$

회귀선 =  $y$ 의 기대값

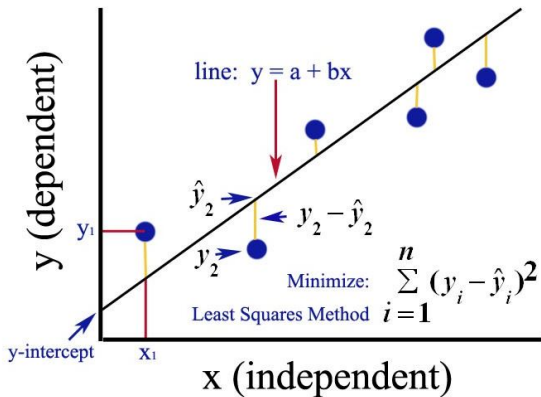
# 회귀계수의 추정

- ▶ 선형 모형이 실제 자료를 얼마나 잘 설명하는지, 즉, 자료를 가지고 선형모형을 적합 하기 위해서는 최소제곱법(method of least square)이라는 방법을 사용한다.
- ▶ 선형회귀모형  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ 에서, 설명변수  $x_1, x_2, \dots, x_n$ 을 선형회귀모형에 대입해서 얻은  $y_i$ 의 추정값과 실제  $y_i$  값의 차이를 구하고, 이 차이들을 제공해서 더한 값을 최소화하는 계수를 구하는 방법을 최소제곱법이라고 부른다.
- ▶ 즉,  $\sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$ 의 값을 최소화하는  $\beta_0, \beta_1$ 의 값을 실제  $\beta_0, \beta_1$ 의 추정값으로 하는 것이다.

# 회귀계수의 추정

어느 직선이 가장 실데이터와 가까울까?

모든 잔차제곱합이 최소가 되는 직선 -> 회귀계수 추정



최소제곱법의 원리

## 회귀계수의 추정 : 예

예 : 앞서 상관분석에 사용된 국어영역과 영어영역의 점수이다.

- ▶ 모국어에 관련된 어학 능력이 외국어에 관련된 어학 능력에 영향을 끼치는지 알아보기 위해, 국어 점수를 설명변수로 하고 영어 점수를 반응변수로 하는 단순회귀모형을 적합해보도록 한다.
- ▶ 회귀계수를 구하는데 사용되는 R Code는 다음과 같다.
- ▶ 적합된 회귀모형은 다음과 같다.

국어 = X, 영어 = Y  
lm(종속~독립)

$$\hat{Y}_i = 15.99 + 0.35 \times x_i$$

해석:  
국어 성적을 1점 올리면  
평균적으로  
영어 성적이 0.35점 오른다.

```
lm(eng~kor)
```

```
Call:
```

```
lm(formula = eng ~ kor)
```

```
Coefficients:
```

(Intercept)	kor
15.99	0.35

# Model formula

- ▶  $eng = \beta_0 + \beta_1 \times kor + \epsilon$ 을 `lm()`에서 사용하는 formula로 표시하면 `eng ~ kor`가 된다.
- ▶ 절편  $\beta_0$ 는 굳이 쓰지 않아도 항상 존재하는 것으로 취급된다.
- ▶ 절편을 제외한 모형  $eng = \beta_1 \times kor + \epsilon$ 을 사용하고 싶다면 `eng~kor-1` 또는 `eng~0+kor`을 사용해야 한다.

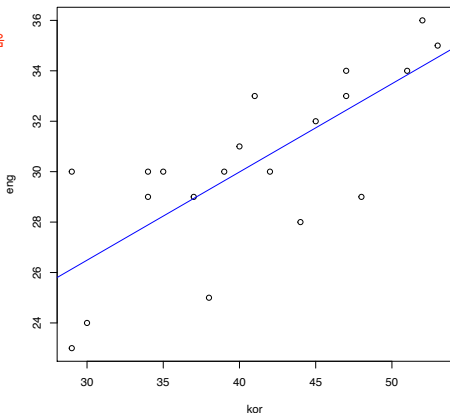
잔차에 대한 가정:  
오차가 정규분포 (평균 0, 분산 시그마제곱) 을 따른다.  
모든 오차항의 분산은 동일하다.  
각 오차항은 독립이다.

잔차분석 = 반드시 해야 됨.

선형회귀분석 결과를 `abline()`에 넘겨주어 바로 그래프에 표시할 수 있다.

```
fit1=lm(eng~kor)
plot(kor,eng)
abline(fit1, col="blue")
```

블루 라인 = Y 기댓값의 추정량  
피티드 라인으로 y를 예측할 수 있음



lm() 으로 찾은 모형을 다음과 같이 살펴볼 수 있다.

```
coef(fit1)      # 회귀 계수

(Intercept)      kor
    15.9899      0.3499

fitted(fit1)[1:6]  # fitted values

    1      2      3      4      5      6
30.69 29.29 33.84 34.54 29.99 28.94

round(residuals(fit1)[1:6],2)  # 잔차

    1      2      3      4      5      6
-0.69 -4.29  0.16  0.46  1.01  0.06

fitted(fit1)[1:6] + residuals(fit1)[1:6]  # eng와 같아야 함

    1  2  3  4  5  6
30 25 34 35 31 29

eng[1:6]

[1] 30 25 34 35 31 29
```

```
confint(fit1)          # 계수의 신뢰구간
```

```
                2.5 %   97.5 %  
(Intercept)  9.7911 22.1886  
kor           0.2002  0.4996
```

베타 0 과 베타 1에 대한 신뢰구간  
0을 포함하고 있지 않다.  
가설 검정을 아직 하지 않았지만, 신뢰 구간을 가지고도  
적당한 결론을 낼 수 있다. 귀무가설을 기각하겠구나..

```
deviance(fit1)        # 잔차제곱합
```

```
[1] 99.03      잔차 = (Y - Y_hat)
```

```
sum((eng - fitted(fit1))^2)  # 잔차제곱합
```

```
[1] 99.03
```



# 예측

lm() 을 통해 만들어진 모델은 predict()를 사용하여 예측할 수 있다.

```
predict(fit1, newdata=data.frame(kor=37))
```

```
1  
28.94
```

```
coef(fit1)[1] + coef(fit1)[2]*37
```

```
(Intercept)  
28.94
```

```
predict(fit1, newdata=data.frame(kor=37), interval="confidence")
```

```
fit   lwr   upr  
1 28.94 27.7 30.17
```

```
# 신뢰구간
```

```
predict(fit1, newdata=data.frame(kor=37), interval="prediction")
```

```
fit   lwr   upr  
1 28.94 23.86 34.02
```

```
# 예측구간
```

# 모형평가

```
summary(fit1)
```

```
Call:
```

```
lm(formula = eng ~ kor)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-4.288	-1.138	0.413	1.613	3.862

잔차에 대한 정보

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	15.9899	2.9505	5.42	3.8e-05
kor	0.3499	0.0713	4.91	0.00011

```
(Intercept) ***
```

```
kor ***
```

```
---
```

```
Signif. codes:
```

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.35 on 18 degrees of freedom
```

```
Multiple R-squared: 0.573, Adjusted R-squared: 0.549
```

```
F-statistic: 24.1 on 1 and 18 DF, p-value: 0.000113
```

절편, 기울기에 대한 추정치  
p 벨류 =  
귀무가설 기각에 대한 근거  
즉 x, y에 대한 관계를 증명하는 수치

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	15.98985	2.95048	5.419	3.78e-05	***
kor	0.34994	0.07125	4.911	0.000113	***

---

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

- ▶ Estimate 열은 절편과 계수의 추정치를 보여줌
- ▶ Pr(> |t|) 열은 p-value로, t 분포를 사용하여 각 변수가 얼마나 유의한지를 알려준다.
- ▶ '\*' 또는 '\*\*\*'로 표시된 문자열은 p-value가 얼마나 유의한지 표시한다. 아무런 표시가 없다면 계수가 통계적으로 유의하지 않음을 뜻한다.

멀티플 R 스퀘어 (결정계수) :  $SSR / SST$   
제곱합분해  $\rightarrow (y_i - \bar{y})$  제곱합 =  $(y_i - \hat{y})$  제곱합 +  $(\hat{y} - \bar{y})$  제곱합  $\rightarrow SST = SSE + SSR$   
등호 오른쪽에는  $x$ 에 대한 정보가 포함됨  
R 스퀘어 = 회귀선의 설명력 (1에 가까워질 수록)

다중회귀분석의 경우 조정결정계수를 확인할 것.

그 의미는 R스퀘어와 같다.

```
Multiple R-squared:  0.5727, Adjusted R-squared:  0.5489  
F-statistic: 24.12 on 1 and 18 DF,  p-value: 0.0001125
```

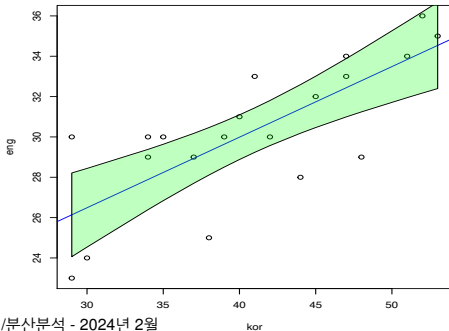
- ▶ 결정계수(Multiple R-squared)는 회귀직선이 얼마나 데이터를 잘 설명하는지에 대한 척도이다. 결정계수가 1이면 모든 점들이 회귀직선 상에 있다.
- ▶ 그러나 결정계수는 설명 변수가 늘어나면 그 값이 커지는 성질이 있으므로 이를 자유도로 나눈 조정결정계수(Adjusted R-squared)가 더 많이 사용된다.
- ▶ F-statistic은 모형에서 기울기가 유의한지에 대한 척도이다. F 분포를 이용해서 p-value가 계산된다. 단순 선형 회귀분석의 경우 앞 슬라이드의 기울기 계수 ( $\beta$ )가 0인지를 검정할때의 p-value와 같다.

$(SSR/\text{자유도}) / (SSE/\text{자유도})$

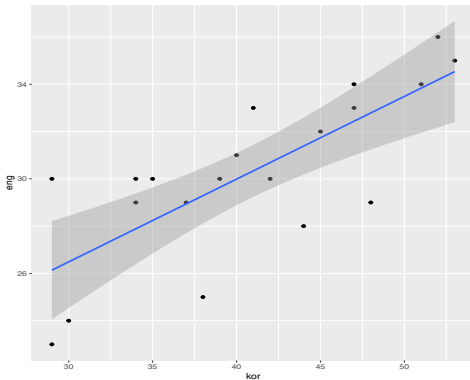
# 신뢰구간을 포함한 회귀 모형 도시하기

```
plot(kor, eng)
sorted.kor=sort(kor, index.return=TRUE)
s.kor=sorted.kor$x
s.eng=eng[sorted.kor$ix]
p <- predict(fit1, interval = "confidence")
s.p <- p[sorted.kor$ix,]
abline(fit1, col = "blue")
x <- c(s.kor, tail(s.kor, 1), rev(s.kor), s.kor[1])
y <- c(s.p[, "lwr"], tail(s.p[, "upr"], 1), rev(s.p[, "upr"]), s.p[, "lwr"])
polygon(x, y, col = rgb(0, 1, 0, 0.25))
```

$E(y)$ 의 신뢰구간  
 $E(y) = \text{베타}_0 + \text{베타}_1 * x$



```
library(ggplot2)
q <- ggplot(data.frame(kor, eng), aes(kor, eng))
q+geom_point()+stat_smooth(method="lm")
```

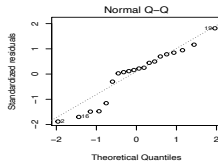
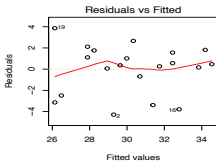


# 모형평가 차트

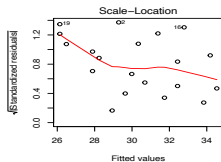
```
par(mfrow=c(2,2))  
plot(fit1)
```

잔차 분석

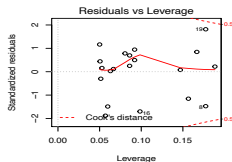
0을 중심으로 골고루 퍼져 있음



위 그림과 같은 정보 담고 있음



잔차에 아웃라이어가 존재하는가?



- ▶ 'Residuals vs Fitted' 차트: 선형 회귀분석에서 오차는 평균이 0이고 분산이 일정한 정규분포임을 가정하였으므로 잔차가 특별한 경향을 보이지 않는 것이 이상적이다.
- ▶ 'Normal Q-Q' 차트: 잔차가 정규분포를 따르는지 확인. 기울기 1인 직선이 되는 것이 이상적이다.
- ▶ 'Scale-Location' 차트: 표준화 잔차가 특별한 경향을 보이지 않는 것이 이상적이다.
- ▶ 'Residuals vs Leverage' 차트: 이상치(Outlier)의 유무를 검사하는데 유용하다.



# 단순회귀분석과 상관계수

```
coef(fit1)[2]

      kor
0.3499

cor(eng, kor) * sd(eng) / sd(kor)

[1] 0.3499

cor(eng, kor)^2

[1] 0.5727

summary(fit1)$r.squared

[1] 0.5727
```

# 직접 해보기

Car 데이터를 가지고 위에서 배운 `lm()`을 이용하여  
 $dist \sim \beta_0 + \beta_1 speed + \epsilon$  모형을 적합하고, 회귀계수 추정, 예측,  
모형평가, `anova`, 신뢰구간 포함한 회귀모형 도식, 모형 평가 차트  
등을 그려보자.

# Car -모형적합

```
m <-lm(dist~speed, data=cars )  
summary(m)
```

Call:

```
lm(formula = dist ~ speed, data = cars)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.07	-9.53	-2.27	9.21	43.20

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-17.579	6.758	-2.60	0.012
speed	3.932	0.416	9.46	1.5e-12

```
(Intercept) *  
speed          ***  
---
```

Signif. codes:

```
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 15.4 on 48 degrees of freedom  
Multiple R-squared: 0.651, Adjusted R-squared: 0.644  
F-statistic: 89.6 on 1 and 48 DF, p-value: 1.49e-12

# Car - 회귀계수 및 잔차

**coef** (m)

(Intercept)	speed
-17.579	3.932

**fitted**(m) [1:6]

1	2	3	4	5	6
-1.849	-1.849	9.948	9.948	13.880	17.813

**residuals** (m) [1:6]

1	2	3	4	5	6
3.849	11.849	-5.948	12.052	2.120	-7.813

**fitted**(m) [1:6] + **residuals** (m) [1:6]

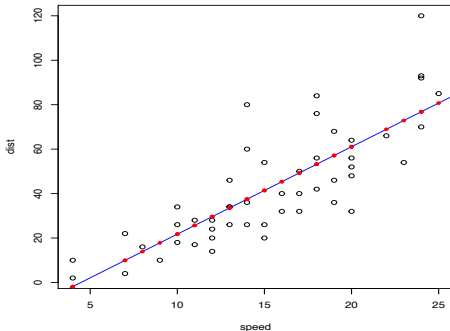
1	2	3	4	5	6
2	10	4	22	16	10

**cars**\$dist [1:6]

[1] 2 10 4 22 16 10

# Car -선형회귀 그래프

```
plot(cars)
abline(m,col='blue')
points(cars$speed,fitted(m),col='red',pch=20)
```



# Car -신뢰구간 및 잔차 제공합

```
confint(m)
```

```
                2.5 % 97.5 %  
(Intercept) -31.168 -3.990  
speed        3.097  4.768
```

```
sum(residuals(m)^2)
```

```
[1] 11354
```

# Car -예측

```
predict(m, newdata=data.frame(speed=3))
```

```
1  
-5.782
```

```
coef(m)[1]+coef(m)[2]*3
```

```
(Intercept)  
-5.782
```

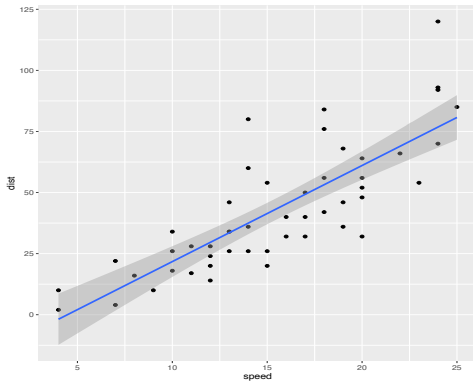
```
predict(m, newdata=data.frame(speed=3), interval='confidence')
```

```
fit    lwr    upr  
1 -5.782 -17.03 5.463
```

```
predict(m, newdata=data.frame(speed=3), interval='prediction')
```

```
fit    lwr    upr  
1 -5.782 -38.69 27.12
```

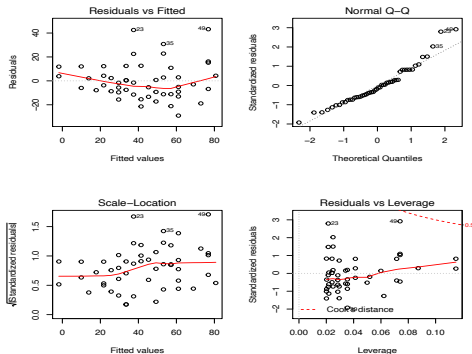
```
q <- ggplot(cars, aes(speed, dist))  
q+geom_point()+ stat_smooth(method='lm')
```





# 모형평가 차트

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



# 범주형 설명변수

## 예 : 붓꽃 data의 선형모형

- ▶ R의 내장 data인 'iris' dataset을 사용한다.
- ▶ 'iris' dataset은 150 송이의 붓꽃들의 꽃받침 길이, 꽃받침 너비, 꽃잎 길이, 꽃잎 너비, 품종을 기록한 dataset이다.
- ▶ 붓꽃의 다른 수치형 변수들(꽃잎 너비, 꽃받침 길이, 꽃잎 길이)이 붓꽃의 꽃받침 너비에 영향을 주는지 알아보기 위한 회귀선형모형을 생각해 볼 수 있다.

- ▶ 범주형 변수를 설명변수로 사용할 경우, 지시변수(Indicator variable)를 사용하여 범주를 표시하게 된다.
- ▶ 지시변수란, 범주에 따라 0과 1 중 하나의 값을 배정하는 정수 형태의 변수로, 범주형 변수를 수치화시켜주는 역할을 한다. 일반적으로, 범주의 갯수가  $p$ 개일 때,  $p-1$ 개의 지시변수를 사용하여 범주형 변수들을 수치화시켜준다.
- ▶ 이러한 지시변수를 이용하여 회귀모형을 적합시켜주면, 각각의 범주에 따라 회귀식의 상수항이 다르게 나타나게 된다.

더미 데이터

Species	지시변수 1	지시변수 2
Setosa	0	0
Versicolor	1	0
Virginica	0	1

붓꽃 자료를 이용하여 만든 지시변수. 2개의 지시변수로 3개의 범주를 구별할 수 있다.

만약 두 지시변수가 둘 다 0이 된다면? 좋은 넓이에 영향을 주지 않는다.

# 범주형 설명변수

R Code를 통해, 붓꽃 자료에 범주형 설명변수인 '품종'을 이용하여 회귀모형을 적합해 보자.

```
iris_lm1 <- lm(Sepal.Width~ Species, data=iris)
summary(iris_lm1)
```

Call:  
lm(formula = Sepal.Width ~ Species, data = iris)

Residuals:

Min	1Q	Median	3Q	Max
-1.128	-0.228	0.026	0.226	0.972

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	3.4280	0.0480	71.36
Speciesversicolor	-0.6580	0.0679	-9.69
Speciesvirginica	-0.4540	0.0679	-6.68

Pr(>|t|)

(Intercept)	< 2e-16 ***
Speciesversicolor	< 2e-16 ***
Speciesvirginica	4.5e-10 ***

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

Residual standard error: 0.34 on 147 degrees of freedom  
Multiple R-squared: 0.401, Adjusted R-squared: 0.393  
F-statistic: 49.2 on 2 and 147 DF, p-value: <2e-16 -> 귀무가설 기각

- ▶ 회귀직선에 대한 F검정과 범주형 변수를 포함한 각각의 회귀계수에 대한 t검정의 결과, 회귀직선 및 모든 회귀계수가 유의하다는 결론을 얻었다.
- ▶ 범주형 변수 'Speciesversicolor'는 'versicolor' 품종에 1을, 나머지 품종에 0을 배정하는 지시변수이다. 또한 'Speciesvirginica'는 1virginica' 품종에 1을, 나머지 품종에 0을 배정하는 지시변수이다.
- ▶ 각각의 지시변수의 계수를 통해, 품종에 따라 상수항이 서로 다른 회귀직선을 얻을 수 있다.

# 분산분석(Analysis of variance, ANOVA)이란?

- ▶ 분산분석이란, 실험계획법(Design of experiments)에서 가장 많이 사용하는 분석방법 중 하나이다.
- ▶ 분산분석은 특성값의 분산, 혹은 변동을 분석하는 방식으로, 특성값의 변동을 제곱합으로 나타내고 이 제곱합을 실험에 관련된 요인의 수준별로 분해하여 오차에 비해 큰 영향을 주는 요인이 무엇인지 찾아내는 분석 방법이다.

요인 설명 변동량 + 설명 불가 변동량 -> 분산분석

## 분산분석을 시행하는 실험의 사례

- ▶ 여러 공법에 의해 생산되는 금속가공품의 인장강도를 비교하기 위해 실험을 한 결과, 특정한 공법에서 강도가 높게 관측되었다고 하자.  
$$\begin{matrix} \text{인장강도} = y(\text{특성값}) \\ \text{공법} = x \end{matrix}$$
- ▶ 이 때, 관측결과는 공법에 따른 강도의 차이로 인해 나타났을 수도 있고, 그 외의 규명되지 않은 요인 (예 : 작업자의 능력) 때문에 나타났을 수도 있다.
- ▶ 공법에 따른 인장강도의 차이가 있는가를 제대로 알아내기 위해서는, 각 공법마다 여러 번의 실험이 이루어져야 한다.
- ▶ 한 공법에 여러 명의 작업자를 할당하여 각각 실험을 하였을 때, 그 공법에서 관측된 자료들의 변동은 작업자들의 능력 차이로 인한 변동으로 생각할 수도 있다. 한편, 서로 다른 공법에서 관측된 인장강도의 평균값이 다르게 나타나는 것은 공법에 따른 변동으로 생각할 수 있다.
- ▶ 따라서, 공법에 따른 변동이 공법 외의 변동보다 크다면, 공법에 따른 인장강도에 차이가 있다고 볼 수 있다.

# 분산분석의 용어

- ▶ 반응변수(특성값) : 관측의 대상이 되는 값(예: 금속공예품의 인장강도) =  $y$
- ▶ 요인(입력변수) : 특성값에 영향을 준다고 판단되는 요소(예: 공법) =  $x$  ( $y$ 값에 영향을 주는 요소)
- ▶ 처리 : 요인의 값 및 서로 다른 요인들의 조합(예: 각각의 서로 다른 공법들) 요인의 수준, 개수, 조합 ...
- ▶ 분산분석 : 자료에서 발생하는 변동성을 모형에 의한 변동(요인 및 처리에 의한 변동)과 오차에 의한 변동(그 외의 규명되지 않은 요인에 의한 변동)으로 분해하고 비교하여 요인의 유의성을 검증하는 분석 방법
- ▶ 요인이 1개 사용되는 분산분석을 일원배치 분산분석, 혹은 일원배치법(One-way ANOVA)이라 하고, 요인이 2개 사용되는 분산분석을 2원배치 분산분석, 혹은 이원배치법(Two-way ANOVA)이라 한다.



# 분산분석의 절차

- (1) 요인과 수준, 그리고 반응변수를 설정한다.
- (2) 실험을 설계한다.(예 : 변인 통제, 실험의 반복수 설정, 실험 순서의 랜덤화)
- (3) 실험을 수행한다.
- (4) 자료를 분석하고 결론을 도출한다.

# 일원배치법

- ▶ 반응변수에 대해서 한 종류의 요인의 영향을 조사하고자 할 때 사용하는 방법이다.
- ▶ 수준이 2개라면? 독립변수에 대한 t-test를 진행할 수 있음
- ▶ 보통, 3개 이상의 처리에 대한 효과를 비교한다. 각 수준에서의 반복 수는 꼭 같을 필요는 없다.
- ▶ 일반적으로, 실험은 랜덤하게 선택된 순서대로 진행된다.  
(완전랜덤화계획)
- ▶ 요인의 수준이 3개, 각각의 수준에 대해 반복수가 5인 실험을 생각해보자. 총  $5 \times 3 = 15$ 개의 실험단위에 대해 랜덤하게 순서를 매겨 먼저 나오는 순서대로 처리를 적용하고 실험을 진행하는 것이다.

# 일원배치법의 자료구조

	처리1	처리2	...	처리 $k$	평균
	$y_{11}$	$y_{21}$	$\cdots$	$y_{k1}$	
	$y_{12}$	$y_{22}$	$\cdots$	$y_{k2}$	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
	$y_{1n_1}$	$y_{2n_2}$	$\cdots$	$y_{kn_k}$	
평균	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	$\cdots$	$\bar{y}_{k\cdot}$	$\bar{y}_{\cdot\cdot}$

$$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}, \quad N = \sum_{i=1}^k n_i$$

일원배치법의 자료구조

- ▶  $\bar{y}_{i\cdot}$  ( $i = 1, 2, \dots, k$ )는 각 처리 별 관측값의 평균이고,  $\bar{y}_{\cdot\cdot}$ 는 전체 관측값의 평균이다. 그리고  $N$ 은 전체 관측값의 갯수(= 실험의 총 횟수)이다.
- ▶ 반복수가 모두 같은 경우,  $n_1 = n_2 = \dots = n_k = n$ ,  $N = kn$ 이 된다.

# 일원배치법 : 모형과 가설

- ▶ 일원배치법에서의 모형은 다음과 같다.  $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ ,  
( $i = 1, \dots, k, j = 1, \dots, n$ ),  $\epsilon_{ij} \sim \text{i.i.d } N(0, \sigma^2)$ .  
설명되는 부분 + 오차  
i → j (수정)
- ▶ 위 모형을 다음과 같이 표현할 수 있다.  $Y_{ij} = \mu_i + \epsilon_{ij}$
- ▶ 일원배치법에서의 귀무가설은 다음과 같다.  $H_0 : \tau_i = 0$  for all  $i$  ( $H_0 : \mu_1 = \dots = \mu_k = \mu$ )

# 일원배치법 : 처리효과의 유의성 검정

- ▶ F 검정통계량의 관측값이  $f_0$ 이면, 유의확률과 기각역은 다음과 같이 주어진다.

유의확률 :  $P(F \geq f_0)$ ,  $F \sim F(k-1, N-k)$

유의수준  $\alpha$ 에서의 기각역 :  $f \geq F_\alpha(k-1, N-k)$

- ▶ 일원배치법에서의 분산분석표는 다음과 같이 나타난다.

요인	제곱합	자유도	평균제곱	F값	유의확률
처리	$SS_{tr}$	$k-1$	$MSS_{tr}$	$MSS_{tr} / MSE$	$P(F > f_0)$
잔차	$SSE$	$N-k$	$MSE$		
계	$SST$	$N-1$			

일원배치법의 분산분석표

# 일원배치법 : 예

## 반복수가 모두 같은 경우

- ▶ 어떤 식물의 가공시 처리액의 농도가 식물의 인장강도에 영향을 미치는지의 여부를 조사하기 위해 처리액의 농도를 각각  $A_1 = 3.0\%$   $A_2 = 3.5\%$   $A_3 = 4.0\%$   $A_4 = 4.5\%$  로 설정하고, 각 처리에 대해 반복수 5회의 처리를 하여 총 20회를 랜덤하게 처리한 후 인장강도를 측정했다.
- ▶ 이 자료에 대해 일원배치법의 모형을 적용하여 처리액의 농도에 따른 인장강도의 차이가 존재하는 지를 알아보고 적절한 가설을 세워 유의수준 5%에서 검정해 보자.

	$A_1$	$A_2$	$A_3$	$A_4$	
	47	51	50	22	
	58	62	38	23	
	51	31	47	28	
	61	46	27	42	
	46	49	23	25	
평균	52.6	47.8	37.0	28.0	총평균 41.35

네 가지 농도에 의한 인장강도

## R Code

```
A1 <- c(47, 58, 51, 61, 46); A2 <- c(51, 62, 31, 46, 49)
A3 <- c(50, 38, 47, 27, 23); A4 <- c(22, 23, 28, 42, 25)
A <- c(A1, A2, A3, A4)
group <- as.factor(rep(1:4, each=5))
fabric <- data.frame(A, group)
A_table <- cbind(A1, A2, A3, A4)
apply(A_table, 2, mean) ; mean(A)
```

y (반응변수)

x (설명변수)

```
      A1      A2      A3      A4
52.6  47.8  37.0  28.0
[1] 41.35
```

```
aov_fabric <- lm(A~group, data=fabric)
anova(aov_fabric)
```

Analysis of Variance Table

```
Response: A
          Df Sum Sq Mean Sq F value Pr(>F)
group      3   1827     609   6.46 0.0045 **
Residuals 16   1508      94
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p 밸류 -> 귀무가설 기각

# 일원배치법 : 예

## 반복수가 일정하지 않은 경우

- ▶ 세 종류의 공정에서 생산된 철선의 인장강도에 차이가 있는지를 알아보기 위해, 각각의 공정에서 인장강도를 측정해보았다. 일원배치법을 이용하여 공정에 따른 인장강도의 차이가 존재하는 지에 대해 유의수준 5%에서 검정해 보자.

	공정1	공정2	공정3	
	2	4	6	
	3	5	5	
	4	6	7	
	5	4	4	
		3	6	
			8	
평균	3.5	4.4	6.0	총평균 4.8

공정에 의한 인장강도



## R Code

```
M1 <- c(2,3,4,5); M2 <- c(4,5,6,4,3); M3 <- c(6,5,7,4,6,8)
M <- c(M1,M2,M3)
group_M <- as.factor(rep(1:3,times=c(4,5,6)))
mean(M1); mean(M2); mean(M3); mean(M)

[1] 3.5
[1] 4.4
[1] 6
[1] 4.8

mechanism <- data.frame(M,group_M)
aov_mechanism <- lm(M~group_M, data=mechanism) lm(설명, 반응변수)
anova(aov_mechanism)

Analysis of Variance Table

Response: M
          Df Sum Sq Mean Sq F value Pr(>F)
group_M    2   16.2    8.10    4.81  0.029 *
Residuals 12   20.2    1.68
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 정리

- ▶ 상관분석을 위해 `cor.test()`를 이용한다.
- ▶ 선형 회귀분석을 위해 `lm()`을 사용한다.
- ▶ 연관된 많은 함수들을 이용하여 찾아낸 모형을 평가할 수 있다: `'confint()'`, `'coef()'`, `'summary()'`, `'anova()'` 등.
- ▶ 범주형 변수의 선형 회귀분석은 각각의 범주에 맞는 회귀모형을 찾는 것으로 생각할 수 있다.
- ▶ 일반적으로, 간단한 모형이 복잡한 모형보다 더 낫다고 간주되고, 모형을 단순화하기 위한 방법으로 ANOVA나 변수 선택을 행한다.
- ▶ `lm()`과 `anova()`를 이용해서 분산분석을 할 수 있다.

## 참고자료

- ▶ Julian J. Faraway, [Linear Models with R](<http://www.amazon.com/Linear-Models-Chapman-Statistical-Science/dp/1439887330>), Second Edition, Chapman Hall/CRC, 2014.
- ▶ 'ggplot2' documentation <<http://docs.ggplot2.org/current/>>
- ▶ 일반통계학, 서울대학교 통계학과, 영지문화사.