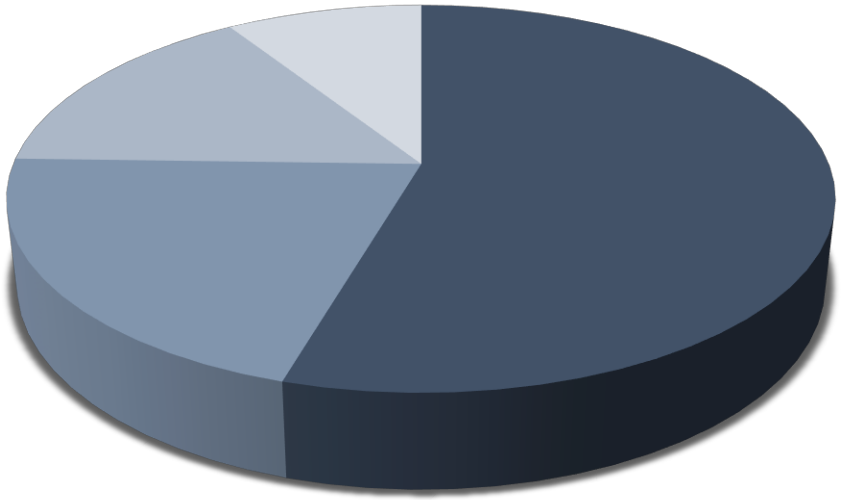
통계학 실습

14. 통계 분석

Contents



Covariance

- 어떤 두 변수에 대한 각각의 평균으로부터 변화하는 방향 및 양에 대해 기대되는 값이다.

- Covariance 값의 부호에 따라 의미가 다르다.
  - + : 한 변수가 증가할 때 다른 변수도 증가하고, 감소할 때 감소한다.
  - - : 한 변수가 증가할 때 다른 변수는 감소하고, 감소할 때 증가한다.

변수 A / 변수 B	증가	감소
증가	+	−
감소	−	+

〈Covariance〉

Covariance 함수 사용

cov(x, y, use, method)

Parameter	option	mean
x, y		두 변수
use	“everything”	NULL 값이 있을 경우, NA로 결과 표시
	“complete.obs”	NULL 값이 있는 경우를 제외하고 표시
method	“pearson”	Pearson correlation coefficient
	“kendall”	Kendall 순위상관계수
	“spearman”	Spearman 순위상관계수

Covariance 예제

- 1. R 분석 예제 데이터 MASS 패키지의 Cars93 데이터 프레임으로 ‘무게’에 따른 ‘마력 ' 의 관계에 대해 분석해보자.

```
> library(MASS)
> str(Cars93)
'data.frame':   93 obs. of  27 variables:
 $ Manufacturer      : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...
 $ Model             : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54 74 73 35 ...
 $ Type              : Factor w/ 6 levels "Compact","Large",..: 4 3 1 3 3 3 2 2 3 2 ...
 $ Min.Price         : num  12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...
 $ Price             : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
 $ Max.Price         : num  18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...
 $ MPG.city          : int  25 18 20 19 22 22 19 16 19 16 ...
 $ MPG.highway       : int  31 25 26 26 30 31 28 25 27 25 ...
 $ AirBags           : Factor w/ 3 levels "Driver & Passenger",..: 3 1 2 1 2 2 2 2 2 2 ...
 $ DriveTrain        : Factor w/ 3 levels "4WD","Front",..: 2 2 2 2 3 2 2 3 2 2 ...
 $ Cylinders         : Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...
 $ EngineSize        : num  1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...
 $ Horsepower        : int  140 200 172 172 208 110 170 180 170 200 ...
 $ RPM               : int  6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...
 $ Rev.per.mile      : int  2890 2335 2280 2535 2545 2565 1570 1320 1690 1510 ...
 $ Man.trans.avail   : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...
 $ Fuel.tank.capacity: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
 $ Passengers        : int  5 5 5 6 4 6 6 6 5 6 ...
 $ Length            : int  177 195 180 193 186 189 200 216 198 206 ...
 $ Wheelbase         : int  102 115 102 106 109 105 111 116 108 114 ...
 $ Width             : int  68 71 67 70 69 69 74 78 73 73 ...
 $ Turn.circle       : int  37 38 37 37 39 41 42 45 41 43 ...
 $ Rear.seat.room    : num  26.5 30 28 31 27 28 30.5 30.5 26.5 35 ...
 $ Luggage.room      : int  11 15 14 17 13 16 17 21 14 18 ...
 $ Weight            : int  2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...
 $ Origin            : Factor w/ 2 levels "USA","non-USA": 2 2 2 2 2 1 1 1 1 1 ...
 $ Make              : Factor w/ 93 levels "Acura Integra",..: 1 2 4 3 5 6 7 9 8 10 ...
>
```

Covariance 예제

- 2. 차가 무겁다면, 엔진이 크고, 힘이 좋아 마력이 더 셀 것이라고 예측할 수 있다. ‘Cov’ 함수를 이용하여 상관 관계에 대한 공분산을 확인하자.

```
> cov(x = Cars93$Weight, y = Cars93$Horsepower, use="everything", method=c("pearson"))
[1] 22825.5
>
```

- Covariance 값이 양수이므로 두 변수는 비례 관계일 것이다.

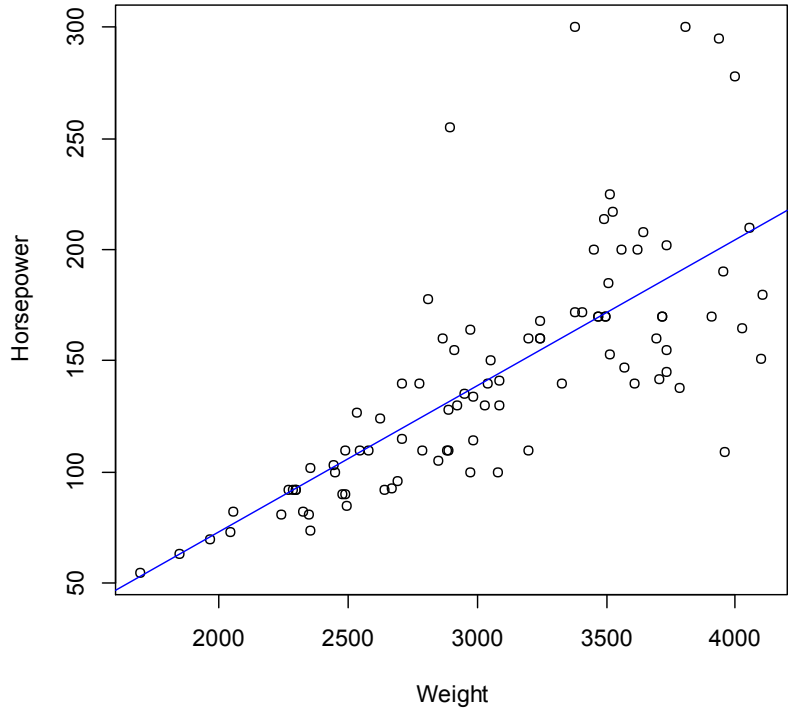
Covariance 예제

- 3. 정말로 두 변수가 비례 관계인지 그래프를 통하여 확인하자.

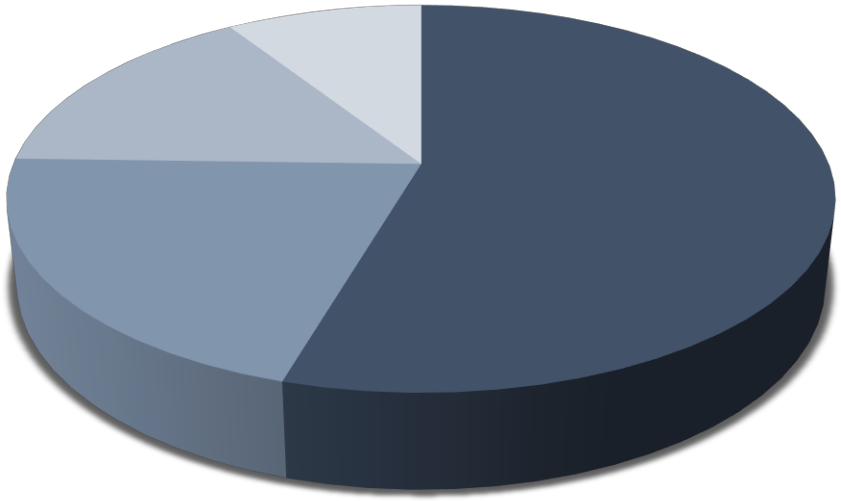
```
> plot(Horsepower ~ Weight, data=Cars93, xlab = "Weight", ylab="Horsepower")
> abline(lm(Horsepower ~ Weight, data=Cars93), col='blue')
>
```

- 그래프를 통하여 보면 무게가 증가할 때, 마력이 증가하는 양상을 볼 수 있다.

- 하지만, 어느 정도로 양의 상관관계를 띄고 있는지는 correlation coefficient 를 통해 알 수 있다.



Contents



Correlation Coefficient

- 상관계수 (Correlation Coefficient)는 표준화된 Covariance이다.

- Correlation Coefficient는 다음과 같은 특징이 존재한다.
  - -1부터 +1까지의 범위를 가진다.
  - +1에 가까울수록 더 강한 양의 선형관계를 가진다.
  - -1에 가까울수록 더 강한 음의 선형관계를 가진다.
  - 0에 가까워질수록 선형관계가 약해지며, 0일 경우 선형관계가 존재하지 않는다.

Correlation Coefficient 예제

- Covariance에 있던 예제를 이용하여 상관계수를 구하여보자.

- cor(x, y, use, method) 함수를 이용하면 된다. use 와 method는 앞서 배운 cov 함수의 인자들과 같은 옵션 및 역할을 가진다.

```
> cor(x = Cars93$Weight, y = Cars93$Horsepower, use="everything", method=c("pearson"))
[1] 0.7387975
> 
```

- 차의 무게와 마력은 약 0.74로 꽤 강한 양의 선형관계를 서로 가지고 있음을 알 수 있다.

Correlation Test

- cor.test(x, y, use, method) 함수를 이용하여 다양한 분석을 확인할 수 있다.
  - p-value를 통한 가설 검정
  - 검정통계량의 값 (t)
  - 95% 신뢰구간
  - 표본상관계수

```
> cor.test(Cars93$Weight, Cars93$Horsepower)

        Pearson's product-moment correlation

data:  Cars93$Weight and Cars93$Horsepower
t = 10.458, df = 91, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6298867 0.8192147
sample estimates:
      cor
0.7387975

> |
```

실습

- 앞서 사용한 MASS 패키지의 Cars93 데이터 프레임에서 차들의 ‘Price’와 해당 차의 ‘Manufacturer’가 어떤 관계인지 1) 그래프를 통해 예상해보고, 2) Covariance와 Correlation Coefficient를 통해 확인해보자.