# 특강, 결과 해석

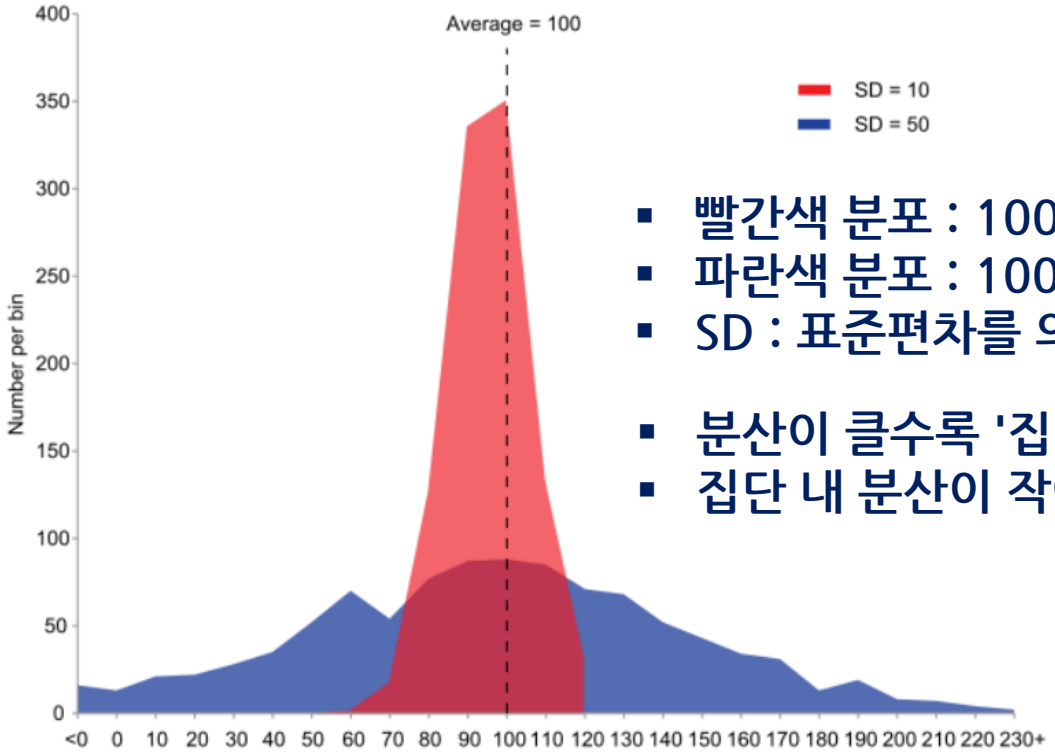
## 2020년, 2021년 기출 관련

# 분산

모집단(population)
모수(parameter)

추출(sampling) →
← 추론(inference)

표본(sample)
통계량(statistic)

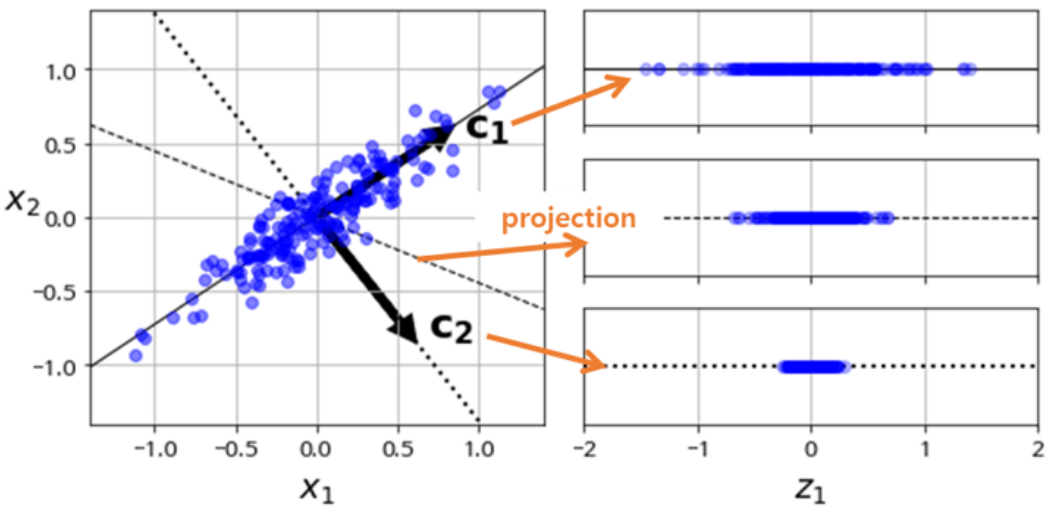


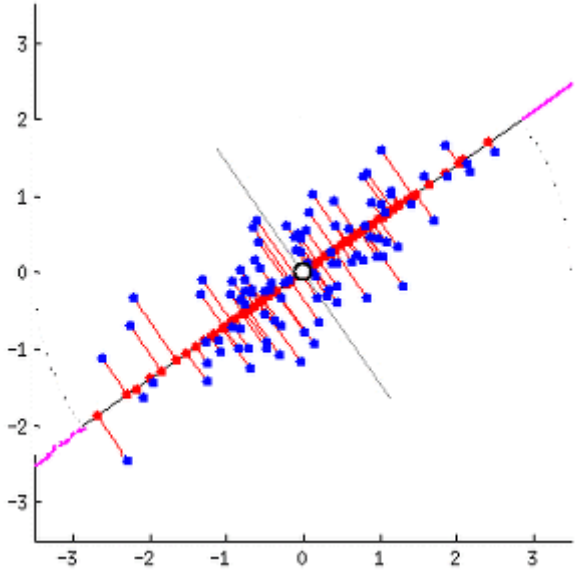

- 빨간색 분포 : 100의 평균값과 100의 분산 값
- 파란색 분포 : 100의 평균값과 2500의 분산 값
- SD : 표준편차를 의미

- 분산이 클수록 '집단의 평균값의 차이'가 무의미해짐
- 집단 내 분산이 작아질수록 평균의 차이가 분명해짐


# PCA(주성분 분석)

- "PCA는 데이터의 분산이 최대가 되는 축을 찾는다" = "정보의 손실을 최소화 한다 "



- 원본 데이터 셋과 투영(projection)된 데이터셋 간의 분산이 최대가 되는 축
  = 평균제곱거리(재구성 오차)를 최소화 하는 축을 찾음
- PCA 좋은 글 : https://laptrinhx.com/dimensionality-reduction-principal-component-analysis-359354885/

# 주성분 분석(PCA) 해석

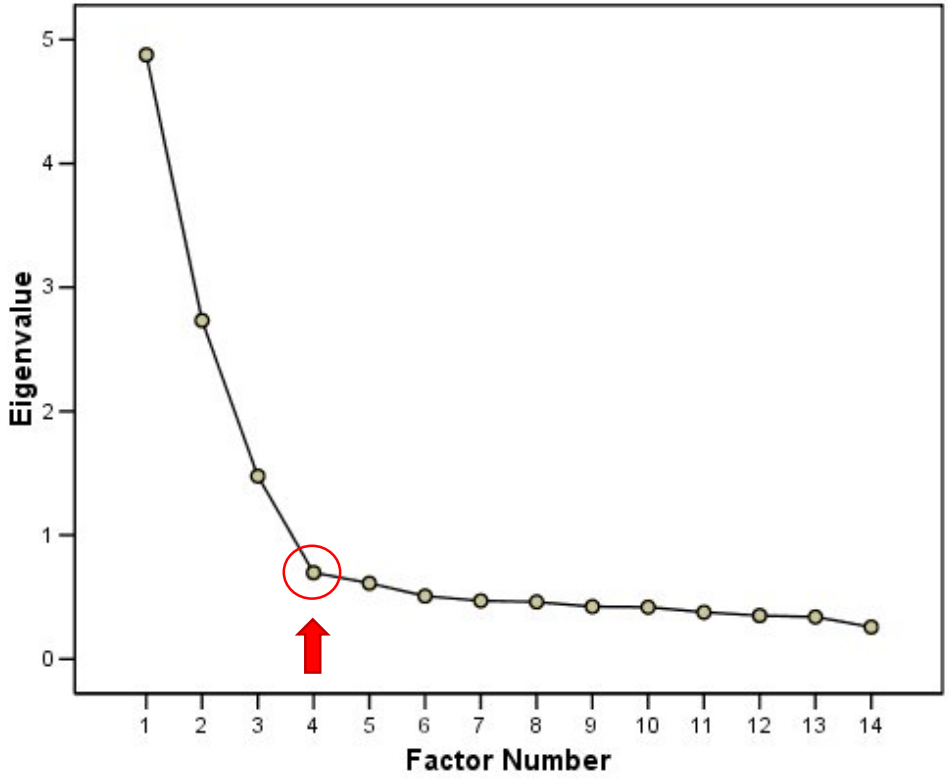
```
> data3 <- princomp(data1, cor=TRUE)   # ISLR 패키지 data (Hitters)
> data3
Call :
princomp(x = data1, cor = TRUE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
2.77339679 2.03026013 1.31485574 0.95454099 0.84109683  0.7237422 0.69841796
생략
17 variables and 263 observations.
> summary(data3)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     2.7733967 2.0302601 1.3148557 0.9575410 0.84109683
Proportion of Variance 0.4524547 0.2424680 0.1016968 0.0539344 0.04161435
Cumulative Proportion  0.4524547 0.6949227 0.7966195 0.8505539 0.89216822
```

- 옆의 결과에서 princomp(data1, cor=TRUE)라는 것이 주성분분석의 함수입니다. cor=TRUE라는 것은 상관계수 행렬을 사용하겠다는 것입니다.
- 이것은 prcomp(data1, scale=TRUE)라는 것과 동일한 동작을 합니다. prcomp의 경우 scale=TRUE라는 것이 상관행렬을 사용하겠다는 것을 의미합니다.
- cor=FALSE, scale=FALSE를 사용하거나 생략하면 공분산 행렬을 사용하겠다는 의미가 됩니다.
- 공분산 행렬은 변수의 측정 단위를 그대로 반영한 것이고,
- 상관행렬을 사용하는 경우 모든 측정 단위를 표준화 한 것입니다.
- 따라서, 공분산 행렬을 이용한 분석의 경우 변수들의 측정 단위에 민감한 특성이 있습니다.

## summary(data3) 에 대한 해석

- Comp.1, Comp.2, ... Comp.5 가 주성분이며 뒤쪽이 생략되어 있습니다.
  (Cumulative Proportion)이 1이 보여야 모든 성분에 대한 내용이 표시된 것입니다.
- Proportion of Variance 가 각 성분의 분산으로 설명력을 의미합니다.
- 주성분은 가장 분산이 높은 것부터 작은 순서로 주성분1, 2, 3 ... 이 됩니다.
- Comp.1 의 경우 0.4524547, Comp.2 의 경우 0.2424680 ... 등으로 표시되고 있습니다
- Cumulative Proportion은 Proportion of Variance를 누적한 것입니다.
  Comp.1은 그대로 표시되고 Comp.2의 경우 Comp.1 + Comp.2의 Proportion of Variance를 더한 것이고,
  Comp.3의 Comp.1 + Comp.2 + Comp.3 의 Proportion of Variance를 더한 것이 됩니다.
- 설명력을 이야기 할 때는 Cumulative Proportion을 보면 됩니다.
- 주성분은 1번부터 순서대로 사용됩니다. 따라서 주성분을 4개 사용한다면 Comp.1에서 Comp.4까지 사용한 것이 됩니다.
- 주성분을 4개 사용했을 때의 설명력은 Comp.4의 Cumulative Proportion을 보면됩니다. (위의 그림에서는 0.8505539가 됩니다. 약 85.05% 입니다.)
- 차원을 2차원으로 줄였다는 것은 2개의 주성분만 사용하겠다는 것입니다.
- 설명력은 전체 주성분을 사용해야 100%가 됩니다.
- 따라서 차원을 2차원으로 줄인 경우 ( 1 - 0.6949227 ) 이 되어서 0.3050773이 되며, 이것을 %로 표현하면 약 30.51% 손실이 됩니다.
- X차원으로 줄였을 때 손실율은 (1 - X차원의 Cumulative Proportion) 입니다.

# 주성분 개수 선택 - Elbow 기법




Scree Plot에서 최적의 요소 수를 찾으라는 문제가 나오면 팔꿈치 부분을 찾아야 합니다.
경사가 완만해지기 시작하는 부분입니다.

## t.test

```
> t.test(x=Default$income, mu=33000)

        One Sample t-test

data:  Default$income
t = 3.8764, df = 9999, p-value = 0.0001067
alternative hypothesis: true mean is not equal to 33000
95 percent confidence interval:
 33255.56 33778.41
sample estimates:
mean of x
 33516.98
```

p-value

- 귀무가설 : income의 평균이 33000과 같다
- 대립가설 : income의 평균이 33000과 같지 않다
- df = 9999,  n = df + 1 = 10000
- 95% 신뢰구간 : 33255.56 ~ 33778.41
- p-value : 0.05보다 작으므로 귀무가설 기각, 대립가설 채택
- x의 평균 (점추정 값) : 33516.98

- 결과해석 - 05 -

## t.test

```
> t.test(x=chickwts$weight, mu=260)

        One Sample t-test

data:  chickwts$weight
t = 0.14137, df = 70, p-value = 0.888
alternative hypothesis: true mean is not equal to 260
95 percent confidence interval:
 242.8301 279.7896
sample estimates:
mean of x
 261.3099
```

p-value

- 귀무가설 : weight의 평균이 260과 같다
- 대립가설 : weight의 평균이 260과 같지 않다
- df = 70 (degree of freedom), n = df + 1 = 71 (관측치 개수)
- 95% 신뢰구간 : 242.8301 ~ 279.7896
- p-value : 0.05보다 큰 값으로 귀무가설을 채택, 대립가설을 기각함
- x의 평균(점추정 값) : 261.3099

- 결과해석 - 06 -

# 다중 선형 회귀

```
> temp <- lm(Fertility~., data=swiss)
> summary(temp)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

데이터셋 : swiss, 종속변수 : Fertility

y절편

독립변수

- 각 변수의 회귀계수의 p-value
- 유의수준 95%에서 0.05 보다 작을 때 유의미
- Examination만 무의미

- df = 41
- n = df + 6(변수개수) = 47

**- 결과해석 - 07 -**

결정계수

수정결정계수

F 통계량의 p-value, 결과의 유의미

# 다중 선형 회귀

```
> summary(Wage)
      year           age                     maritl           race           education     
 Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480   1. < HS Grad      :268  
 1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293   2. HS Grad        :971  
 Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190   3. Some College   :650  
 Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37   4. College Grad   :685  
 3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55                   5. Advanced Degree:426  
 Max.   :2009   Max.   :80.00                                                                   
```

education은 범주형 변수

- 결과해석 - 08 -

## 다중 선형 회귀

```
> result <- lm(wage ~ education, Wage)
> summary(result)

Call:
lm(formula = wage ~ education, data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max
-112.31  -19.94   -3.09   15.33  222.56

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 84.104      2.231  37.695  < 2e-16 ***
education2. HS Grad         11.679      2.520   4.634 3.74e-06 ***
education3. Some College    23.651      2.652   8.920  < 2e-16 ***
education4. College Grad    40.323      2.632  15.322  < 2e-16 ***
education5. Advanced Degree 66.813      2.848  23.462  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.53 on 2995 degrees of freedom
Multiple R-squared:  0.2348,    Adjusted R-squared:  0.2338
F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

데이터셋 : Wage, 종속변수 : wage, 독립변수 : education

회귀계수 : 종속변수 wage와의 관계를 나타내는 값

y절편

독립변수

회귀식의 모든 변수가 통계적으로 유의미

education에 대한 더미 변수
더미 변수 개수 = 범주개수 - 1

n = 2995 + 5

- 결과해석 - 09 -

# 다중 선형 회귀

```
> summary(lm(Salary ~., data=Hitters))

Call:
lm(formula = Salary ~ ., data = Hitters

Residuals:
    Min      1Q  Median      3Q     Max
-907.62 -178.35  -31.11  139.09 1877.04
```

Division이 범주형 변수이기 때문에
더미변수(dummy)로 만들어져 사용됨
DivisionW 일 때 1, DivisionE 일 때 0

- DivisionW의 Estimate가 음수이기 때문에 E인 선수에 비해 W인 선수가 평균적으로 Salary가 낮게 됨
- 만일, DivisionW가 양수였다면 E인 선수에 비해 평균적으로 Salary가 높게 됨

- lm을 이용해서 선형 회귀 분석을 했는데, Multiple R-squared 가 0.5461로 점수가 매우 낮음
- 변수들의 회귀계수에 대한 p-value를 보았을 때 0.05보다 큰 것이 많음
- 따라서 선형인지 아닌지 알 수 없음

- 결과해석 -



범주형

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  163.10359   90.77854   1.797 0.073622 .
AtBat         -1.97987    0.63398  -3.123 0.002008 **
Hits           7.50077    2.37753   3.155 0.001808 **
HmRun          4.33088    6.20145   0.698 0.485616
Runs          -2.37621    2.98076  -0.797 0.426122
RBI           -1.04496    2.60088  -0.402 0.688204
Walks          6.23129    1.82850   3.408 0.000766 ***
Years         -3.48905   12.41219  -0.281 0.778874
CAtBat        -0.17134    0.13524  -1.267 0.206380
CHits          0.13399    0.67455   0.199 0.842713
CHmRun        -0.17286    1.61724  -0.107 0.914967
CRuns          1.45430    0.75046   1.938 0.053795 .
CRBI           0.80771    0.69262   1.166 0.244691
CWalks        -0.81157    0.32808  -2.474 0.014057 *
LeagueN       62.59942   79.26140   0.790 0.430424
DivisionW   -116.84925   40.36695  -2.895 0.004141 **
PutOuts        0.28189    0.07744   3.640 0.000333 ***
Assists        0.37107    0.22120   1.678 0.094723 .
Errors        -3.36076    4.39163  -0.765 0.444857
NewLeagueN   -24.76233   79.00263  -0.313 0.754218
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 315.6 on 243 degrees of freedom
  (결측으로 인하여 59개의 관측치가 삭제되었습니다.)
Multiple R-squared:  0.5461,    Adjusted R-squared:  0.5106
F-statistic: 15.39 on 19 and 243 DF,  p-value: < 2.2e-16
```

# 다중 선형 회귀 - 변수 선택

```
> model <- lm(Salary~., data=Hitters)
> step(model, direction='backward')
Start:  AIC=3046.02
Salary ~ AtBat + Hits + HmRun + Runs + RBI + Walks + Years +
    CAtBat + CHits + CHmRun + CRuns + CRBI + CWalks + League +
    Division + PutOuts + Assists + Errors + NewLeague

             Df Sum of Sq      RSS    AIC
 - CHmRun     1      1138 24201837 3044.0
 - CHits      1      3930 24204629 3044.1
 - Years      1      7869 24208569 3044.1
 - NewLeague  1      9784 24210484 3044.1
 - RBI        1     16076 24216776 3044.2
 - HmRun      1     48572 24249272 3044.6
 - Errors     1     58324 24259023 3044.7
 - League     1     62121 24262821 3044.7
 - Runs       1     63291 24263990 3044.7
 - CRBI       1    135439 24336138 3045.5
 - CAtBat     1    159864 24360564 3045.8
 <none>                   24200700 3046.0
 - Assists    1    280263 24480963 3047.1
 - CRuns      1    374007 24574707 3048.1
 - CWalks     1    609408 24810108 3050.6
 - Division   1    834491 25035190 3052.9
 - AtBat      1    971288 25171987 3054.4
 - Hits       1    991242 25191941 3054.6
 - Walks      1   1156606 25357305 3056.3
 - PutOuts    1   1319628 25520328 3058.0
```



- direction='backward' 이므로 후진 제거법
- 후진 제거법은 모든 설명변수가 포함된 모형에서 시작
- 한 번 제거된 변수는 다시 모형에 포함될 수 없음
- 변수 선택에 있어 AIC가 작을 수록 좋은 평가이므로, AIC가 작아지게 되는 변수를 제거하게 된다

- AIC가 작을 수록 좋은 평가
- Start :AIC=3046.02 인데 아래 변수명과 AIC 의 목록에서 AIC는 해당 변수를 제거했을 때 AIC가 어떻게 변한다를 표현한 것입니다.
- 그러므로 AIC가 가장 작아지게 되는 변수를 제거하여 더 작은 값을 갖도록 만든다는 의미입니다.

# 로지스틱 회귀

```
> model = glm(default ~ ., data=Default, family=binomial)
> summary(model)

Call:
glm(formula = default ~ ., family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

로지스틱 회귀 (분류)

종속변수 : default, 2항분류

범주형

n = 10000

student가 범주형 변수이기 때문에
더미변수(dummy)로 만들어져 사용됨
studentYes 일 때 1, studentNo 일 때 0

student 값이 Yes 일 때, default를 감소 시킴

$$default = -1.087e+01 - 6.468e-01 * studentYes + 5.737e-03 * balance + 3.033e-06 * income$$

student값이 No 일 때, default를 변화시키지 않음

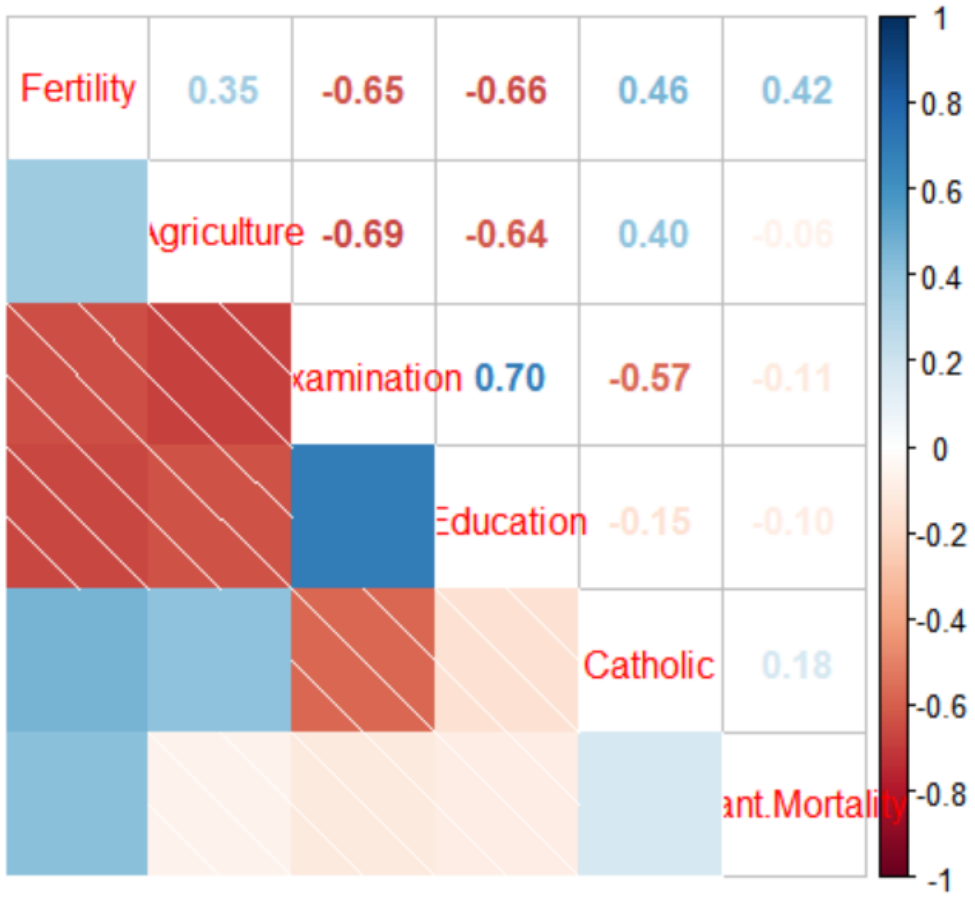
$$default = -1.087e+01 + 5.737e-03 * balance + 3.033e-06 * income$$

- 결과해석 - 12 -

## 상관 계수

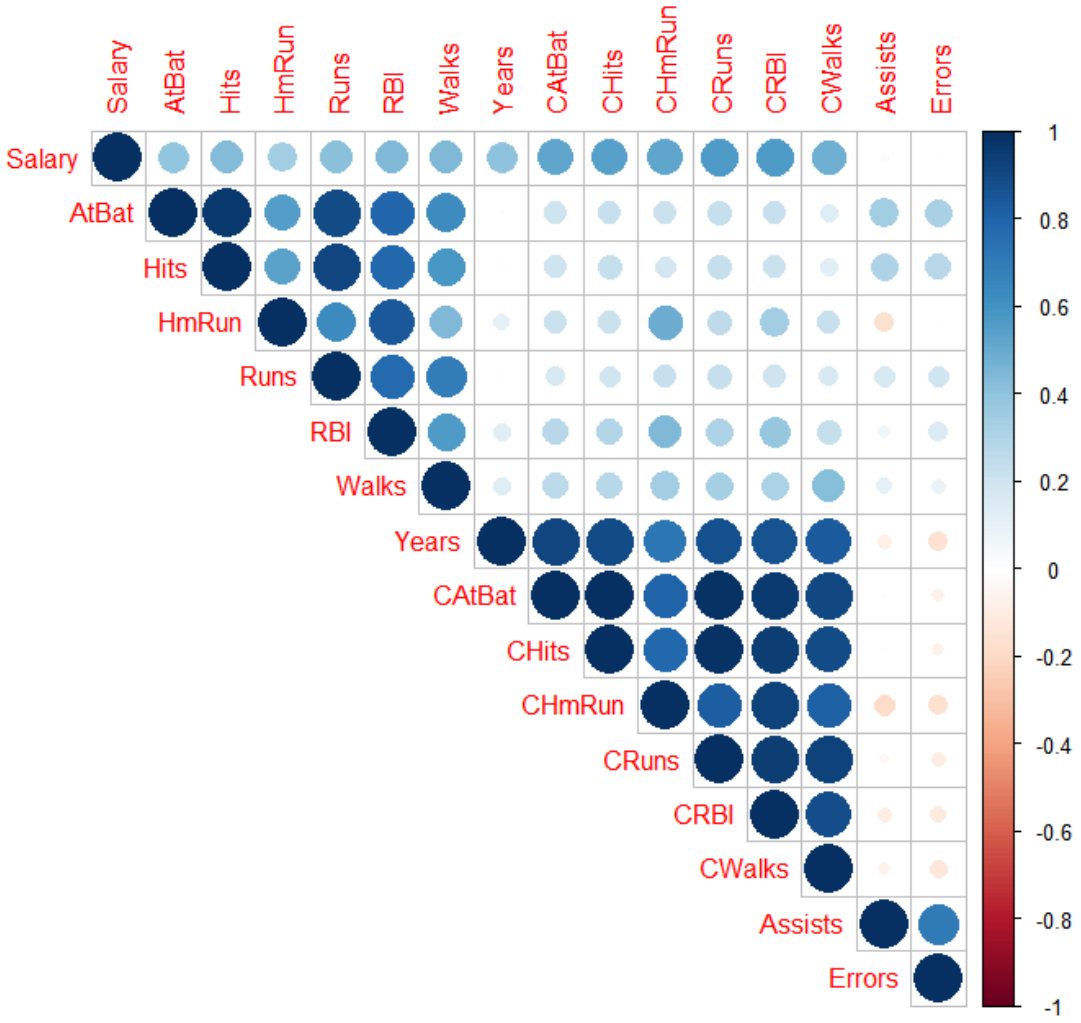
- **상관 계수** 값의 범위는 −1부터 +1까지입니다.
- **계수**의 절대값이 클수록 변수 사이에 강한 **관계**가 있습니다.
- **피어슨 상관**의 경우 절대값 1은 완전한 **선형 관계**를 나타냅니다.
- 0에 가까운 **상관** 값은 변수 사이에 **선형의 상관 관계**가 없음을 나타냅니다.
- 상관계수가 0이라는 것은 아무런 관계가 없다는 것이 아니라 , '**선형의 상관 관계가 아니다**'선로 해석

# 상관계수 그래프 (corrplot)




- 시험(Examination)과 가장 상관관계가 높은 변수는 교육수준(Education)이다
- 교육수준(Education)이 높을수록 출산율(Fertility)은 낮아진다
- 출산율(Fertility)과 농업인구 비율(Agriculture)은 선형관계를 보인다
- 틀림 : 출산율(Fertility)은 시험(Examination)과 가장 높은 음의 상관관계를 가진다

# 상관계수 그래프 (corrplot)




동그라미의 크기가 크고, 짙은 색상일 수록 높은 상관관계

- Salary와의 상관계수가 작은 변수 중 하나는 Errors이다
- Salary와 Errors의 산점도에서는 선형성이 나타나지 않을 것이다.
- Salary를 종속변수로 나머지 변수들을 독립변수로 하는 회귀모형을 적합할 때 다중공선성이 존재할 가능성이 크다
- 틀림:Salary와 CRuns의 상관계수는 통계적으로 유의하다

Salary, Errors 의 상관계수를 보면 거의 흰색으로 보이지 않습니다. 즉, 0에 가깝다는 것이며, 이런 경우 선형성이 없다고 판단할 수 있습니다.

여러 변수들이 진하고 검은 동그라미로 색칠된 관계 (-1 또는 1) 인 것을 볼 수 있습니다.
이런 경우 다중 공선성이 존재한다고 할 수 있습니다.

# 카이제곱 독립성 검정

```
> table(Default$default, Default$student)

default    No  Yes  student
    No   6850 2817
    Yes   206  127
> chisq.test(Default$default, Default$student)

        Pearson's Chi-squared test with Yates' continuity correction

data:  Default$default and Default$student
X-squared = 12.117, df = 1, p-value = 0.0004997
```

**default(연체)에 대한 학생, 비학생의 독립성 검정**

**p-value가 0.05보다 작으므로 귀무가설 기각 대립가설 채택**
**➔ 연체와 학생은 서로 독립이 아니다.**
**➔ 연체와 학생의 차이가 5% 유의 수준에서 존재한다!**

## 카이제곱 검정 (Chi square test)

- 범주형 자료로 구성된 데이터 분석에 사용
- 관찰된 빈도가 기대되는 빈도와 유의한 차이가 있는지 검증
- 카이제곱값($\chi^2$) = $\sum$(관측값 – 기댓값)$^2$/기댓값

## 카이제곱 검정 중 독립성 검정

- Contingency table에 있는 두 개 이상의 변수가 서로 독립인지 검정

- 귀무가설 : 두 변수는 차이가 없음 (독립O, 관계X)
- 대립가설 : 두 변수는 차이가 있음 (독립X, 관계O)

**귀무가설 : 연체와 학생은 서로 독립이다.**
**대립가설 : 연체와 학생은 서로 독립이 아니다.**

## 회귀모형의 anova

Cars 데이터에서 속도(speed)와 제동거리(dist)의 관계를 회귀모형으로 추정한 것이다.
(회귀모형의 유의성 분석)

```
> out <- lm(dist~speed, data=cars)
> anova(out)
Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value

MSE = 오차 분산의 불편추정량

- 회귀계수는 5% 수준에서 유의하다
- 관측치는 48 + 2 이다 (Residuals df + 변수 2개)
- 결정계수 = SSR/SST= ((21186) / (21186 +11354)) = 0.651
- 오차 분산의 불편추정량은 'MSE' 오차제곱평균으로, Mean Sq 와 Residuals가 교차되는 지점에 235.6 이라고 써 있음