Chapter 4 : Resampling

Newton's three sisters

March 2, 2025

Department of Statistics
Sungshin Women's University

Outline

1 Cross-Validation

Cross-Validation

- Assuming that k is an integer that divides N, $1/k$ of data are used for the test, and the other $1 - k/1$ of the data are used to estimate the model.

	Group 1	Group 2	$\cdots$	Group $k-1$	Group k
First	Test	Estimate	$\cdots$	Estimate	Estimate
Second	Estimate	Test	$\cdots$	Estimate	Estimate
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(k-1)$th	Estimate	Estimate	$\cdots$	Test	Estimate
kth	Estimate	Estimate	$\cdots$	Estimate	Test

Example with R code

- making linear regression

```r
cv.linear= function(X, y, k){
  n = length(y)
  m = n/k # k needs to divide n
  S = 0
  for (j in 1:k){
    test = ((j-1)*m+1):(j*m)
    # specify which out of the n pairs are used for test
    beta = solve(t(X[-test,])%*%X[-test,])%*%t(X[-test,])%*%y[-test]
    # estimate beta using data other than those used for test
    e = y[test] - X[test,]%*%beta
    S = S + drop(t(e)%*%e)
    # evaluate the coefficient beta using data for test
  }
  return(S/n)
}
```

Apply data set

```
# Data generation
n = 100
p = 5
X = matrix(rnorm(n*p), ncol = p)
X = cbind(1, X)
beta = rnorm(p + 1)
beta[c(2, 3)] = 0
eps = rnorm(n)
y = X%*%beta + eps

# Evaluation via Cross - Validation
cv.linear(X[, c(1, 4, 5, 6)], y, 10)
```

```
## [1] 0.971863
```

```
cv.linear(X[, c(1, 2, 3, 4)], y, 10)
```

```
## [1] 6.347372
```

```
cv.linear(X, y, 10)
```

```
## [1] 1.004586
```

Drawing plot