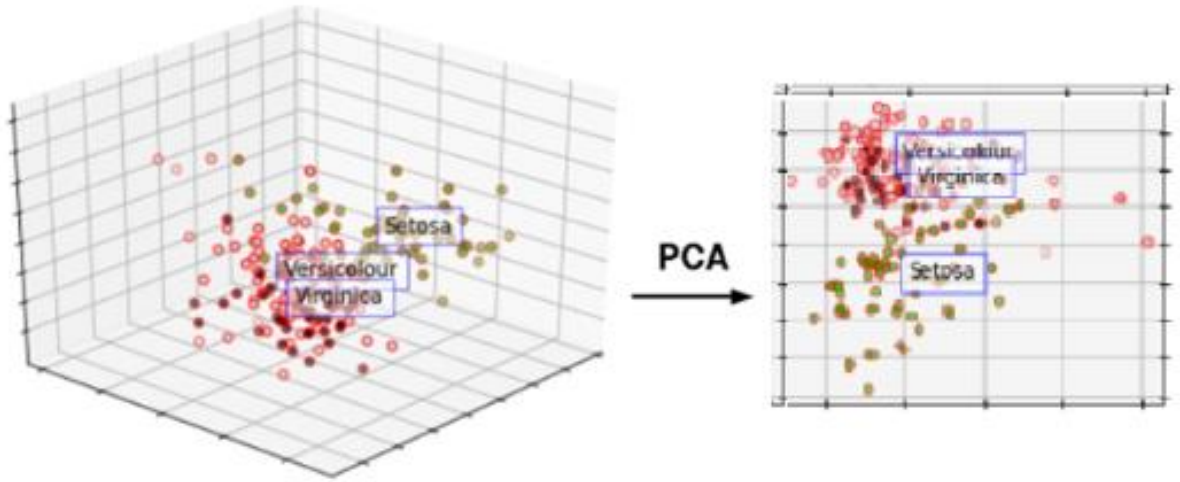
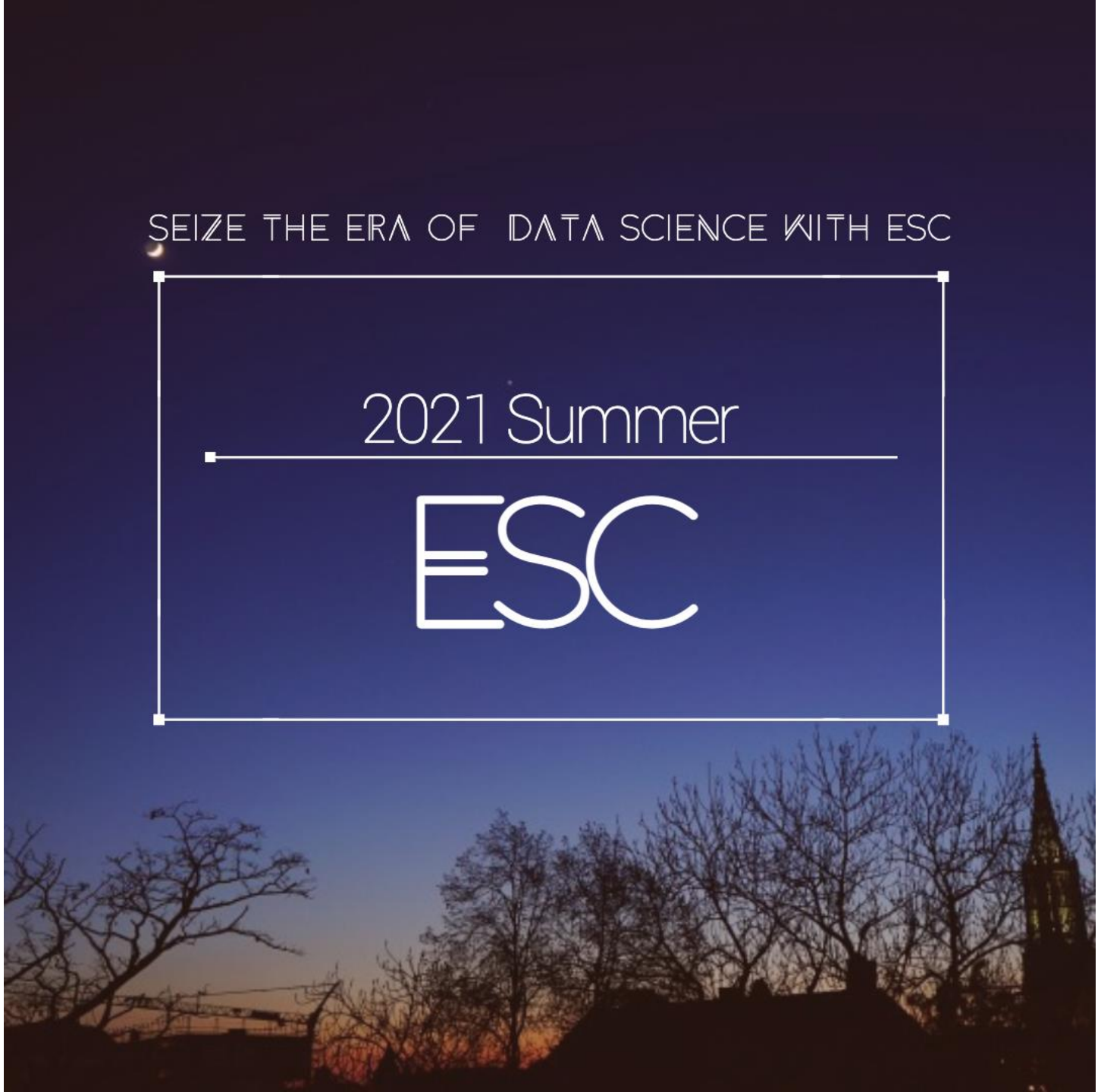
Dimension Reduction

차원 축소

이규민



ESC



차원 축소란?

그럼 2학기에는?

선형 회귀

ESC

21 Summer 학술주제

21 Summer Topic:

High Dimensional Reduction

TextBook : An Introduction to Statistical Learning(G. James, 외) 외 두 권

Recommendation:

선형대수 및 미분적분학, 수리통계학, 회귀분석
통계 관련 지식
R 및 Python 활용 능력

Goal:

회귀분석을 기반으로 한 차원축소 방법론들을 배운 후,
이어지는 2학기에 머신러닝을 체계적으로 배워 활용하고자 한다.

ESC

차원 축소란?

그럼 2학기에는?

선형 회귀

Dimension Reduction

: the process of reducing the dimension of your feature set

Table 5.1 Parameter estimates in Main effects model

Parameter	Estimate	SE
Intercept	-9.273	3.838
Color(1)	1.609	0.936
Color(2)	1.506	0.567
Color(3)	1.120	0.593
Spine(1)	-0.400	0.503
Spine(2)	-0.496	0.629
Weight	0.826	0.704
Width	0.263	0.195

Table 5.2 Backward (Color:4,Spine:3,Width)

Model	Prediction	Model df	Deviance	df	AIC	Models Compared	Deviance Difference
0	Saturated	173	0				
0a	CSW	24					
1	$CS + CW + SW$	18	173.7	155	209.7		
2	$C + S + W$	7	186.6	166	200.6	(2)-(1)	12.9(df=11)
3a	$C + S$	6	208.8	167	220.8	(3a)-(2)	22.2(df=1)
3b	$S + W$	4	194.4	169	202.4	(3b)-(2)	7.8(df=3)
3c	$C + W$	5	187.5	168	197.5	(3c)-(2)	0.9(df=2)
4a	C	4	212.1	169	220.1	(4a)-(3c)	24.6(df=1)
4b	W	2	194.5	171	198.5	(4b)-(3c)	7.0(df=3)
5	$(C = dark) + W$	3	188.0	170	194.0	(5)-(3c)	0.5(df=2)
6	None (Constant)	1	225.8	172	227.8	(6)-(5)	37.8(df=1)

ESC

차원 축소란?

그럼 2학기에는?

선형 회귀

Dimension Reduction

: the process of reducing the dimension of your feature set

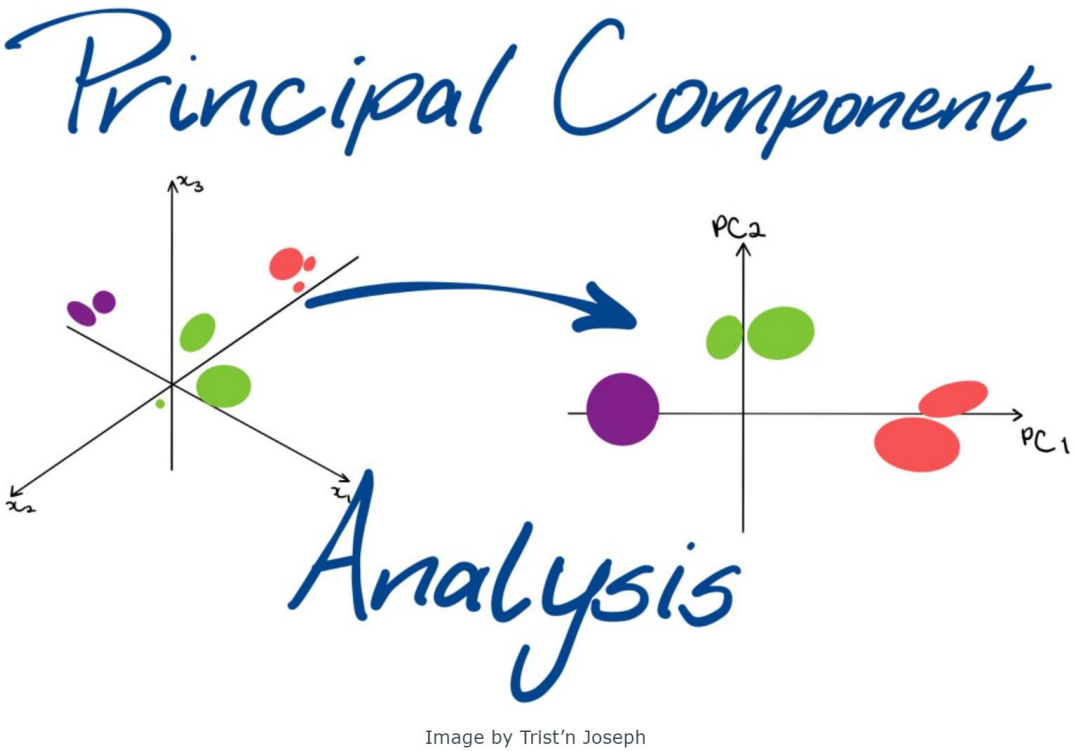
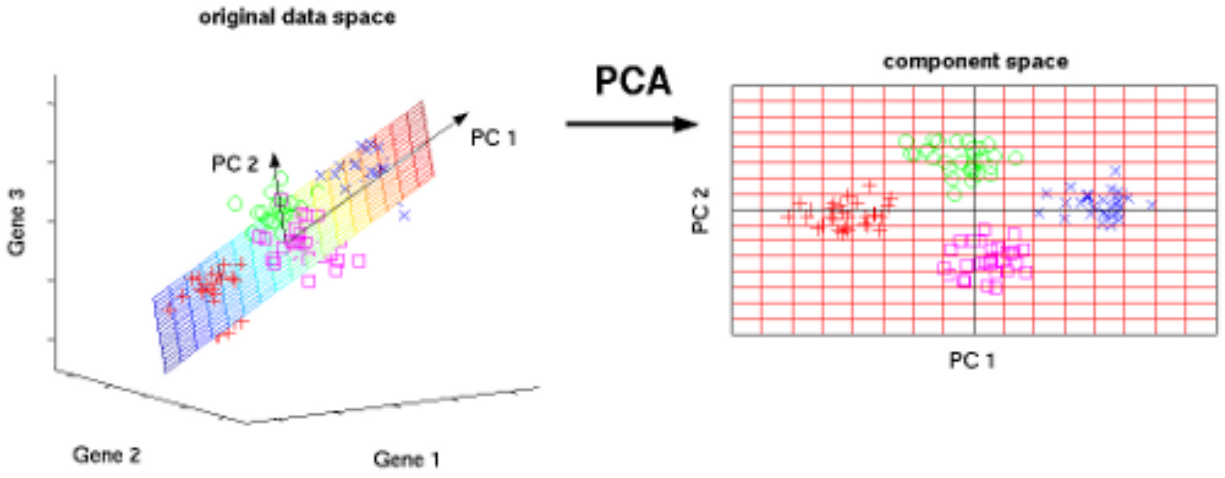


Image by Trist'n Joseph



ESC

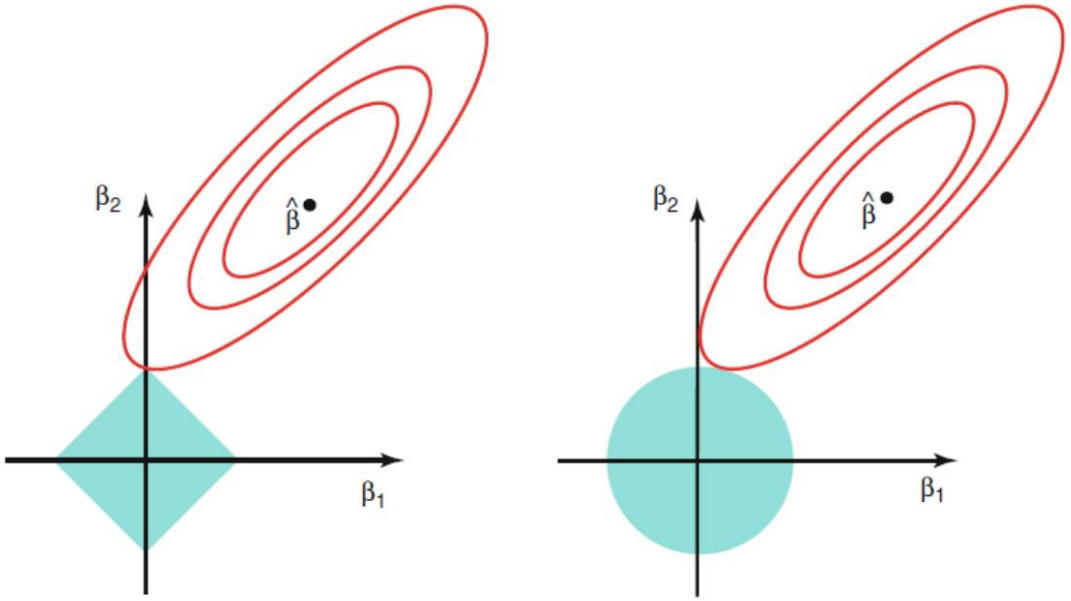
차원 축소란?

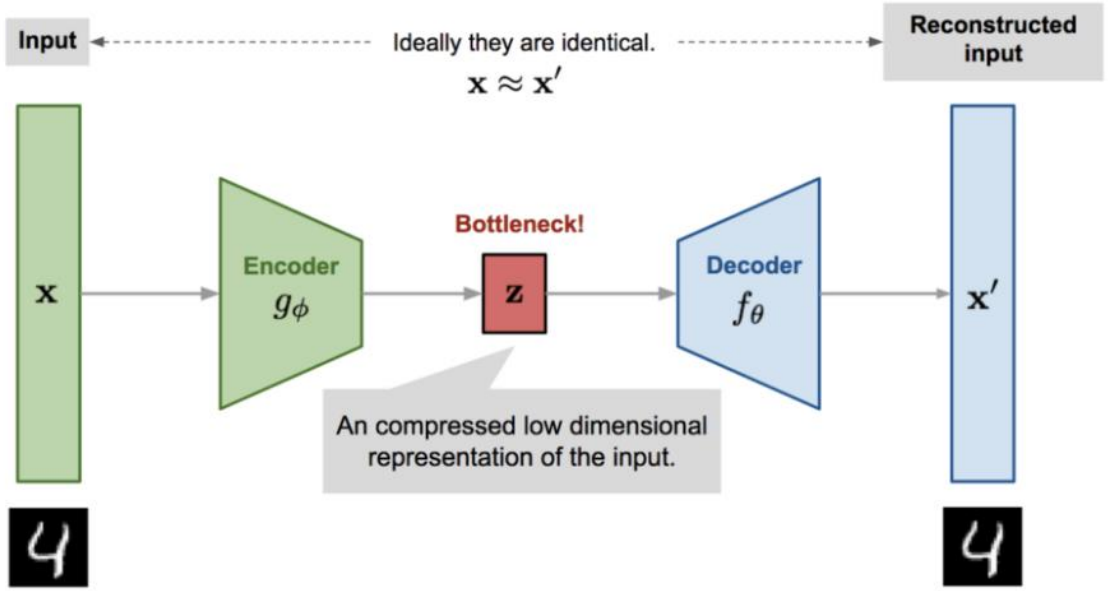
그럼 2학기에는?

선형 회귀

Dimension Reduction

: the process of reducing the dimension of your feature set

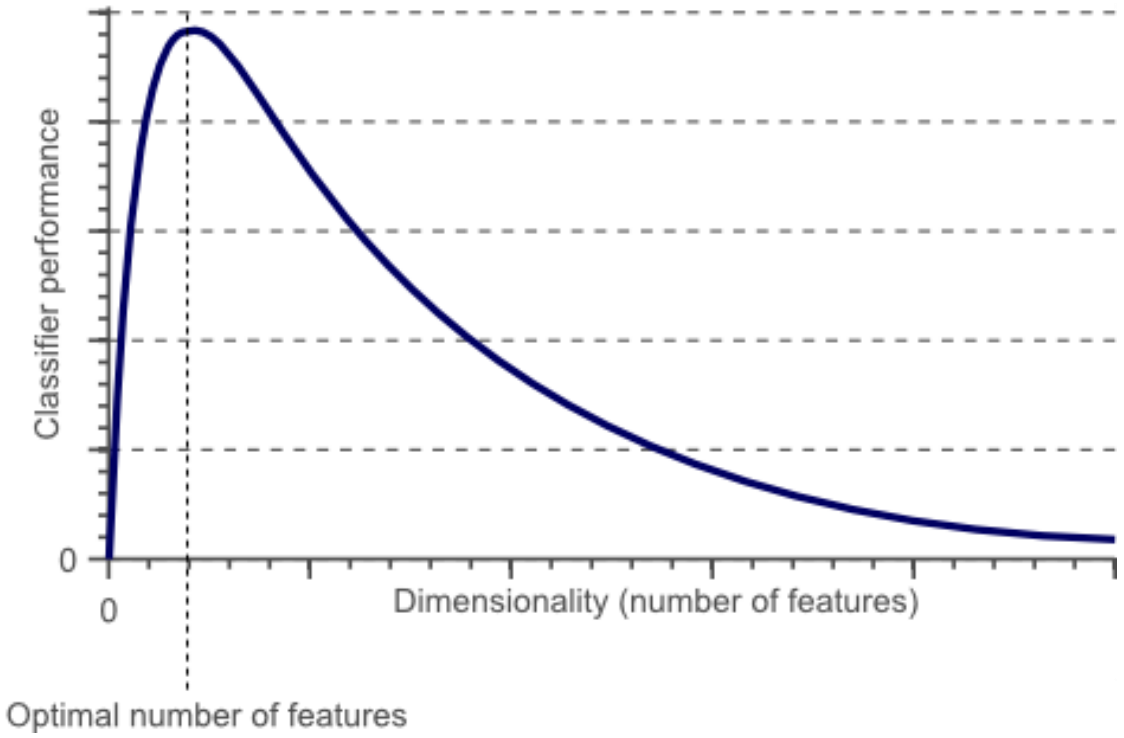


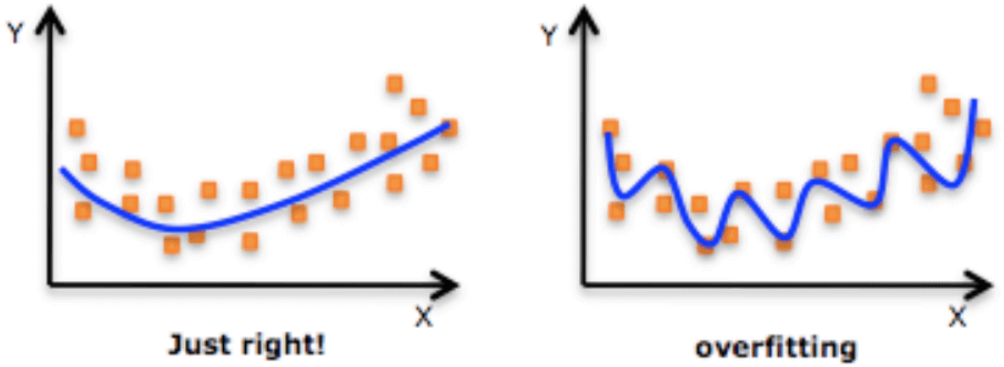


ESC

Why?

Avoids the curse of dimensionality





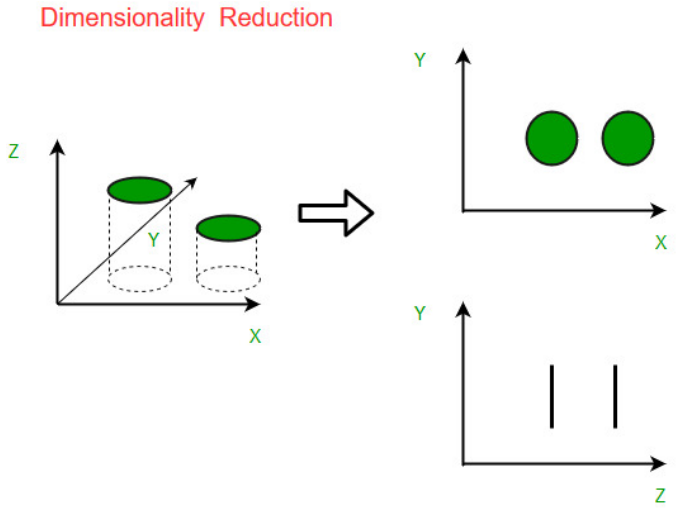
ESC

Why?

Removes Multicollinearity(when predictor variables are highly correlated)
: improves the interpretation of the parameters

Computational issues
: less computing → faster training!
: less data → less storage needed!

Easier to visualize



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-152.9983	68.47747	-2.23	0.0264 *
Weight	-0.380969	0.190594	-2.00	0.0467 *
Height	1.7995873	0.982061	1.83	0.0681
BMI	31.511739	9.293593	3.39	0.0008 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	76.780999	10.04121	7.65	<.0001 *
Weight	0.263259	0.015363	17.14	<.0001 *
Height	-1.488292	0.158734	-9.38	<.0001 *

ESC

차원 축소란?

그럼 2학기에는?

선형 회귀

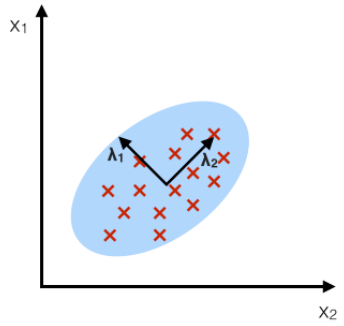
Feature Selection

: Likelihood based methods (AIC, BIC, …)
: Statistical tests (ANOVA, chi square test, …)
: Variance threshold

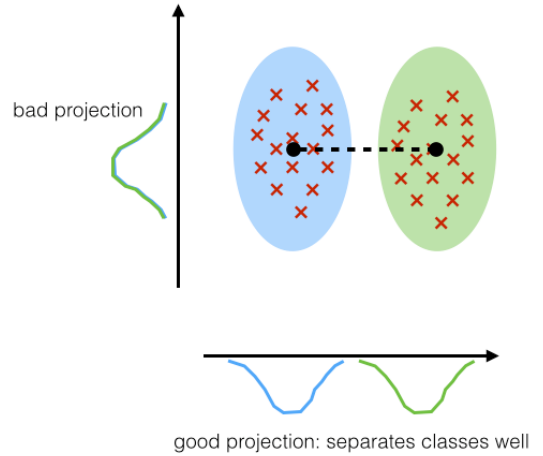
Feature Extraction

: Principal Component Analysis
: Factor Analysis
: Linear Discriminant Analysis

PCA:
component axes that maximize the variance

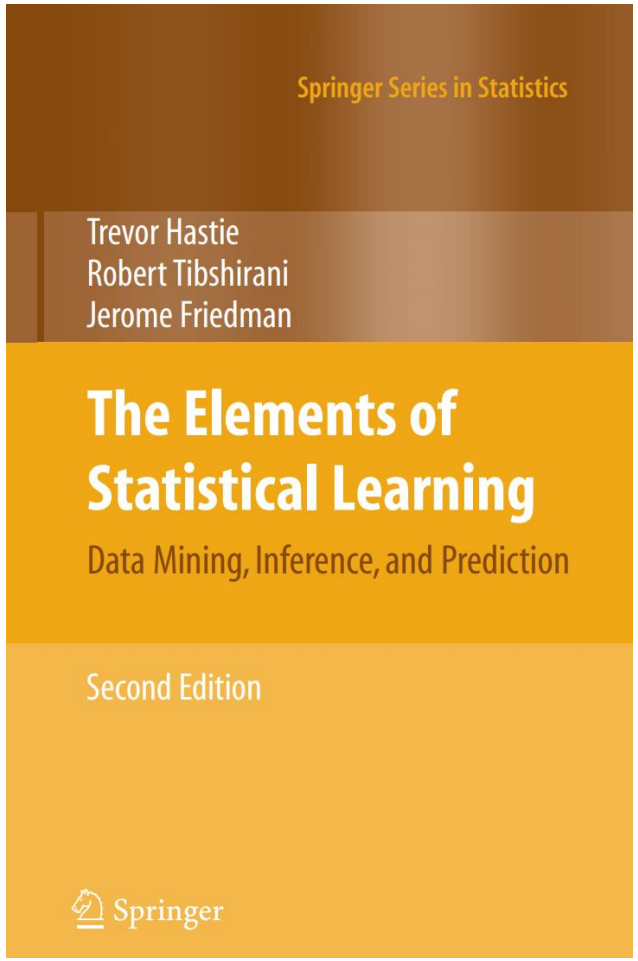


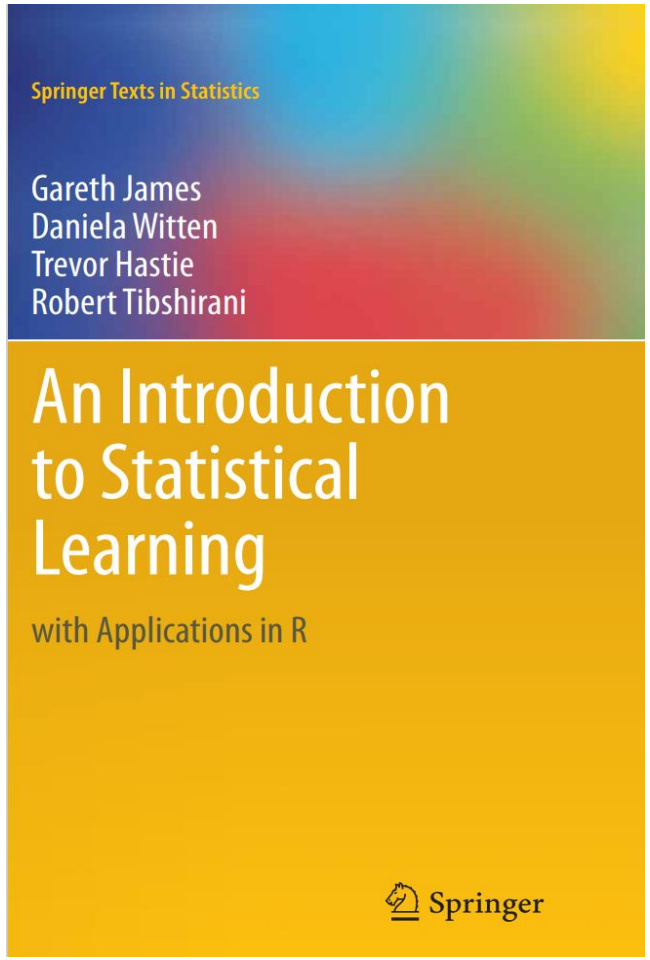
LDA:
maximizing the component axes for class-separation



ESC

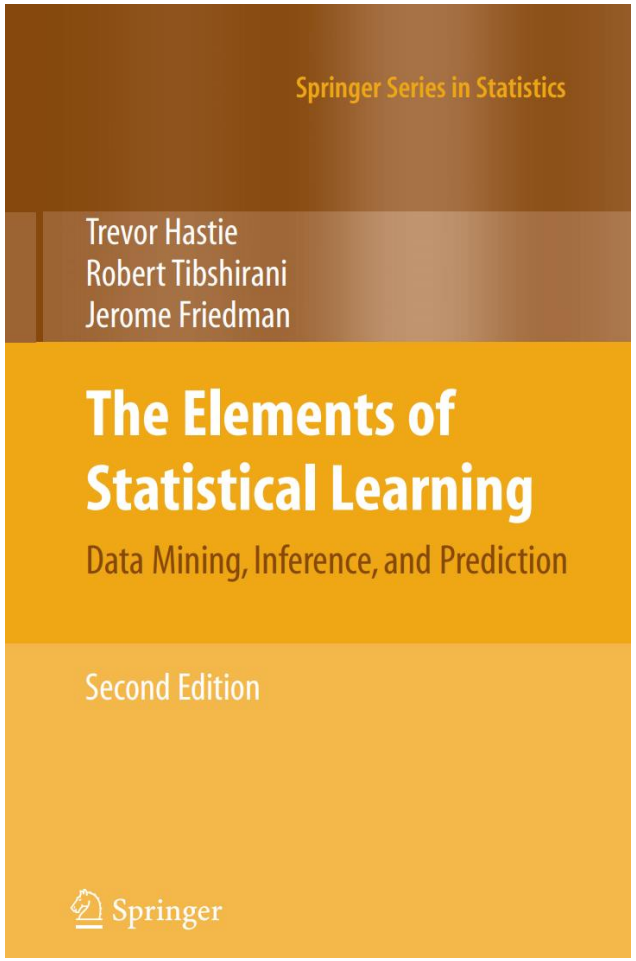
그럼 2학기에는……?


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

ESC

그럼 2학기에는……?



4. Classification
5. Basis expansion and regularization
6. Kernel smoothing methods
7. Model assessment
8. Model inference and averaging
9. Additive models, trees, and related methods
10. Boosting and additive trees
12. Support vector machines and flexible discriminants
14. Unsupervised learning
15. Random forest
16. Ensemble learning

중 일부!

ESC

Linear Regression

선형 회귀

이규민

ESC

차원 축소란?

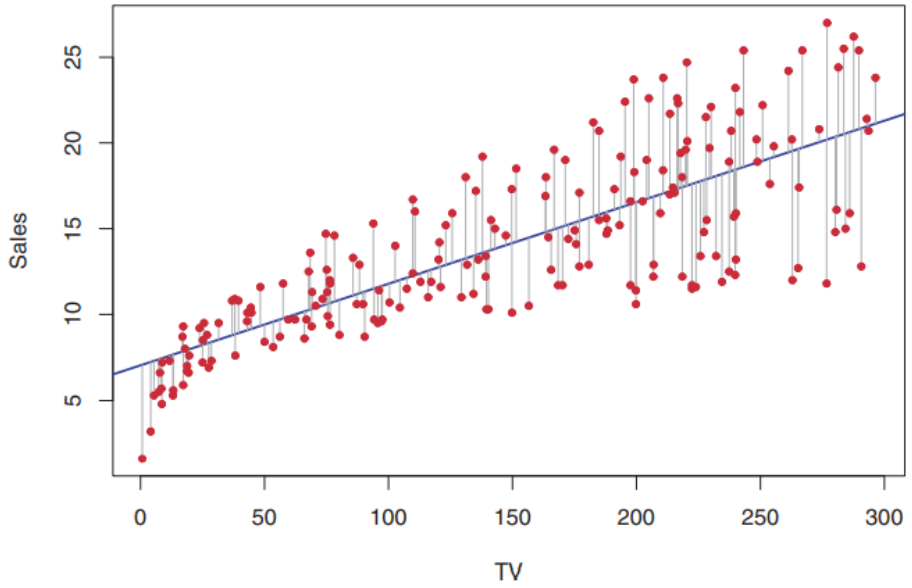
그럼 2학기에는?

선형 회귀

Linear Regression

: want to find relation between Y(dependent variable, response variable) and X(independent variable, explanatory variable, predictor)

: 회귀 계수들이 linear한 것!



$$Y = 0.0475X + 7.03$$

$$Y = f(X) + \epsilon$$

$$= \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \sim iid\ N(0, \sigma^2)$$

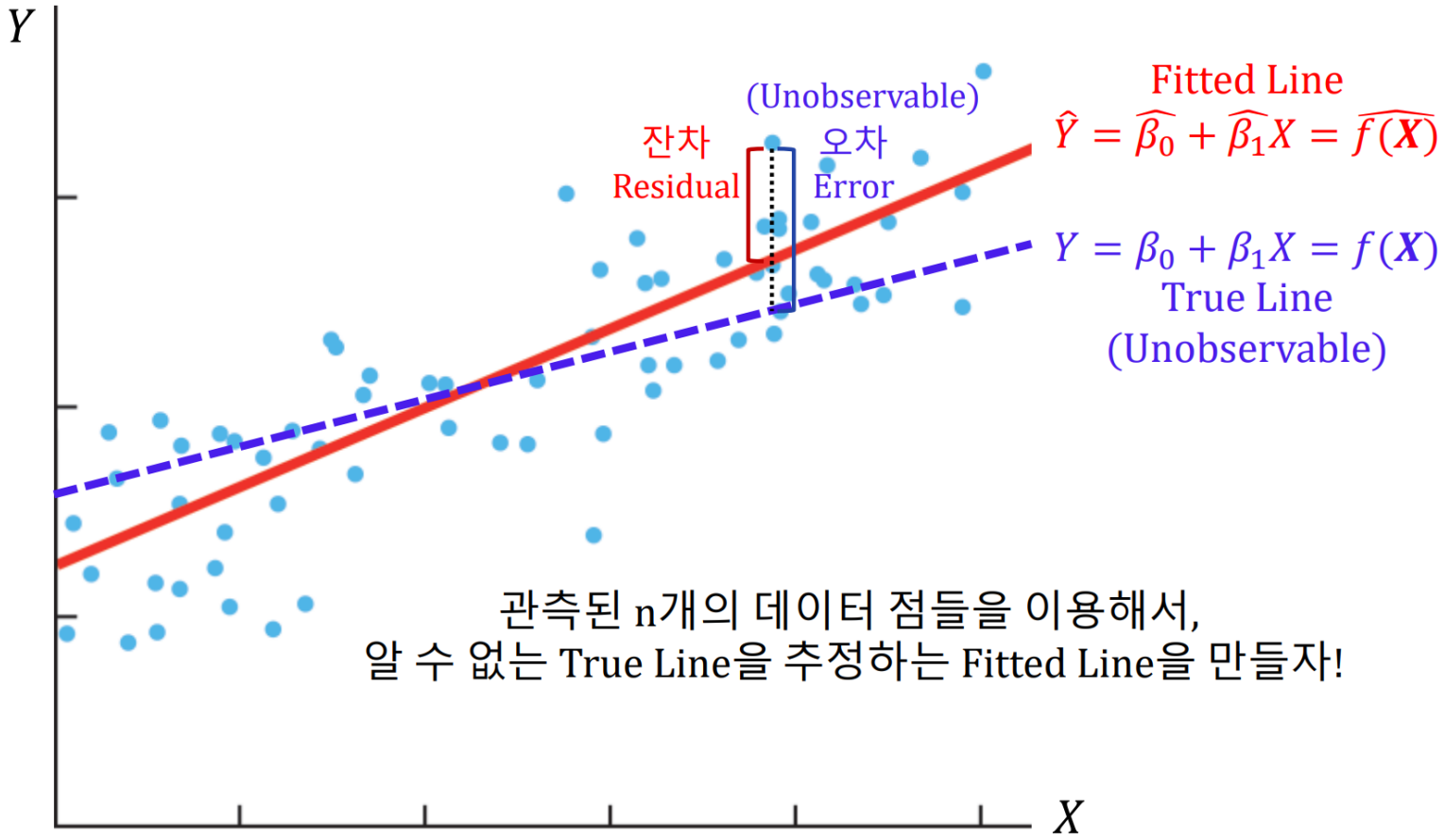
ESC

True line
: 알 수 없다.....

$$Y = \beta_0 + \beta_1 X$$

Fitted line
: data를 이용하여 true line을 추정한 것

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$



관측된 n개의 데이터 점들을 이용해서,
알 수 없는 True Line을 추정하는 Fitted Line을 만들자!

ESC

Estimating coefficients

1. Least Square Estimator
: 잔차 제곱 합을 최소화하는 coefficient를 구하자!

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\frac{\partial Q}{\partial \beta_0} = 0, \frac{\partial Q}{\partial \beta_1} = 0$ 을 만족하는 $\beta_0 = b_0, \beta_1 = b_1$ 을 구하면

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$b_1 = S_{xy} / S_{xx}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

ESC

Estimating coefficients

2. Maximum Likelihood Estimator
: (log) likelihood를 최대화하는 coefficient를 구하자!

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$을 바탕으로 pdf를 구하면

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2}(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma})^2)$$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i)$$

$$logL(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^{n} logf(y_i) = \frac{-n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

LSE와 같은 방법으로 미분해서 0이 되는 값을 구하면 된다!

$$\Rightarrow \quad b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad , \quad b_0 = \bar{y} - b_1 \bar{x}$$

ESC

Properties of coefficients (Gauss-Markov thm)

1. Linear in Y

2. Unbiased estimators

3. Minimum variance (later……)

⇒ Best Linear Unbiased Estimator(BLUE)

What about σ^2?

SSE (Error Sum of Squares)

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

MSE (Error Mean Square)

$$\sum_{i=1}^{n} e_i^2/(n-2) = SSE/(n-2)$$

$\hat{\sigma^2}$ = MSE

※ By MLE,

$$\hat{\sigma^2_{MLE}} = SSE/n = \frac{n-2}{n}\hat{\sigma^2}$$

ESC

Inference

$$E(b_1) = \beta_1$$

$$Var(b_1) = \sigma^2 / S_{xx}$$

$$s^2(b_1) = \hat{\sigma^2} / S_{xx}$$

$$SSE/\sigma^2 \sim \chi^2(n-2)$$

$$\downarrow$$

$$\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

$$\frac{b_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$$

Confidence interval

$$b_1 \pm s(b_1) \times t(1-\alpha/2, n-2)$$

Hypotheses test

$$\rightarrow \quad H_0 : \beta_1 = 0 \quad \text{test statistic} \quad t^* = \frac{b_1}{\hat{\sigma}\sqrt{(1/n + \bar{X}^2/S_{xx})}} \sim t(n-2)$$

$$H_1 : \beta_1 \neq 0 \quad \text{reject if} \quad t^* > t(1-\alpha/2, n-2)$$

ESC

Inference

$$E(b_0) = \beta_0$$

$$Var(b_0) = \sigma^2(1/n + \bar{X}^2/S_{xx})$$

$$s^2(b_0) = \hat{\sigma^2}(1/n + \bar{X}^2/S_{xx})$$

$$SSE/\sigma^2 \sim \chi^2(n-2)$$

↓

$$\frac{b_0-\beta_0}{\sigma\sqrt{(1/n+\bar{X}^2/S_{xx})}} \sim N(0,1)$$

$$\frac{b_1-\beta_1}{\hat{\sigma}\sqrt{(1/n+\bar{X}^2/S_{xx})}} \sim t(n-2)$$

→

$$b_0 \pm s(b_0) \times t(1-\alpha/2, n-2)$$

and tests also……

ESC

ANOVA approach

SSTO
: total deviance $\sum (Y_i - \bar{Y})^2$

SSE
: deviance of true value from fitted value $\sum (Y_i - \hat{Y}_i)^2$

SSR
: deviation of fitted value from mean $\sum (\hat{Y}_i - \bar{Y})^2$

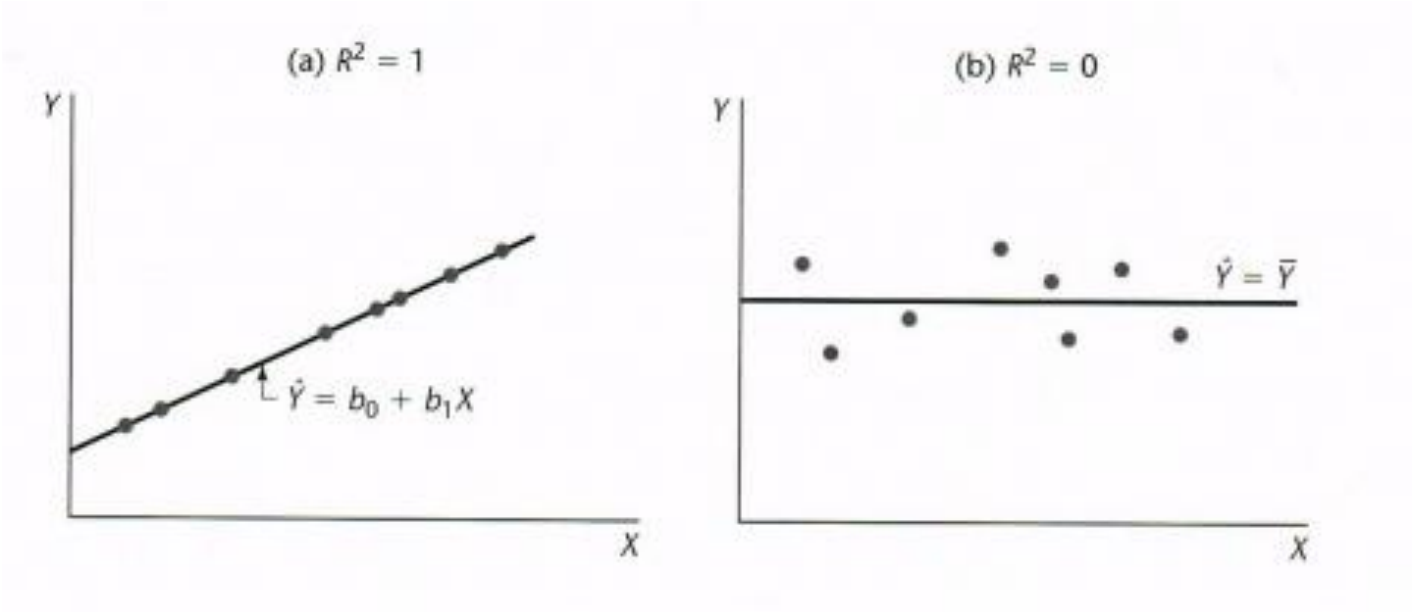
$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

ESC

ANOVA approach

SSTO = SSE + SSR 을 이용해 Coefficient of determination (R^2) 정의

$$R^2 = SSR/SSTO = 1 - SSE/SSTO$$



ANOVA approach

MSE

$\sum_{i=1}^{n} e_i^2/(n-2) = SSE/(n-2)$

MSR

$SSR/1$

ANOVA table:

Source of Variation	Sum of Squares (SS)	df	Mean Square (MS)	E{MS}
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

$H_0 : \beta_1 = 0$ test statistic $F^* = MSR/MSE \sim F(1, n-2)$ under H_0

$H_1 : \beta_1 \neq 0$ reject H_0 if $F^* > F(1-\alpha, 1, n-2)$

ESC

차원 축소란?

그럼 2학기에는?

선형 회귀

그래서 이번 방학에는?

ESC-21 SUMMER 커리큘럼			
WEEK	날짜	Session contents	참고 자료
1	7/8	OT (주제 소개)	ISL 3.1,
2	7/15	Linear Regression	ISL 3.2, 3.3 ESL 3.1, 3.2
3	7/22	Variable Selection	ISL 3.3, 6.1 ESL 3.3
4	7/29	Multicollinearity & Ridge, LASSO	ISL 6.2 ESL 3.4
5	8/5	Principal Component Analysis	ISL 6.3 ESL 3.5 AMSA 11
6	8/12	Factor Analysis	ESL 14.7 AMSA 12
7	8/19	Linear Discriminant Analysis	ISL 4.4 ESL 4.3 AMSA 13
		한 주 쉬고.....!	
1	9/2	21-2 가을 첫 세션 시작!	

* ISL, ESL, AMSA는 각각 "An Introduction to Statistical Learning(G. James, 외)", "The Elements of Statistical Learning(T. Hastie 외)", 그리고 "Applied Multivariate Statistical Analysis(W. Hardle 외)"입니다.

* 빅콘 대회 참여 후 토요일 격주 세션을 진행할 것입니다.

ESC

감사합니다