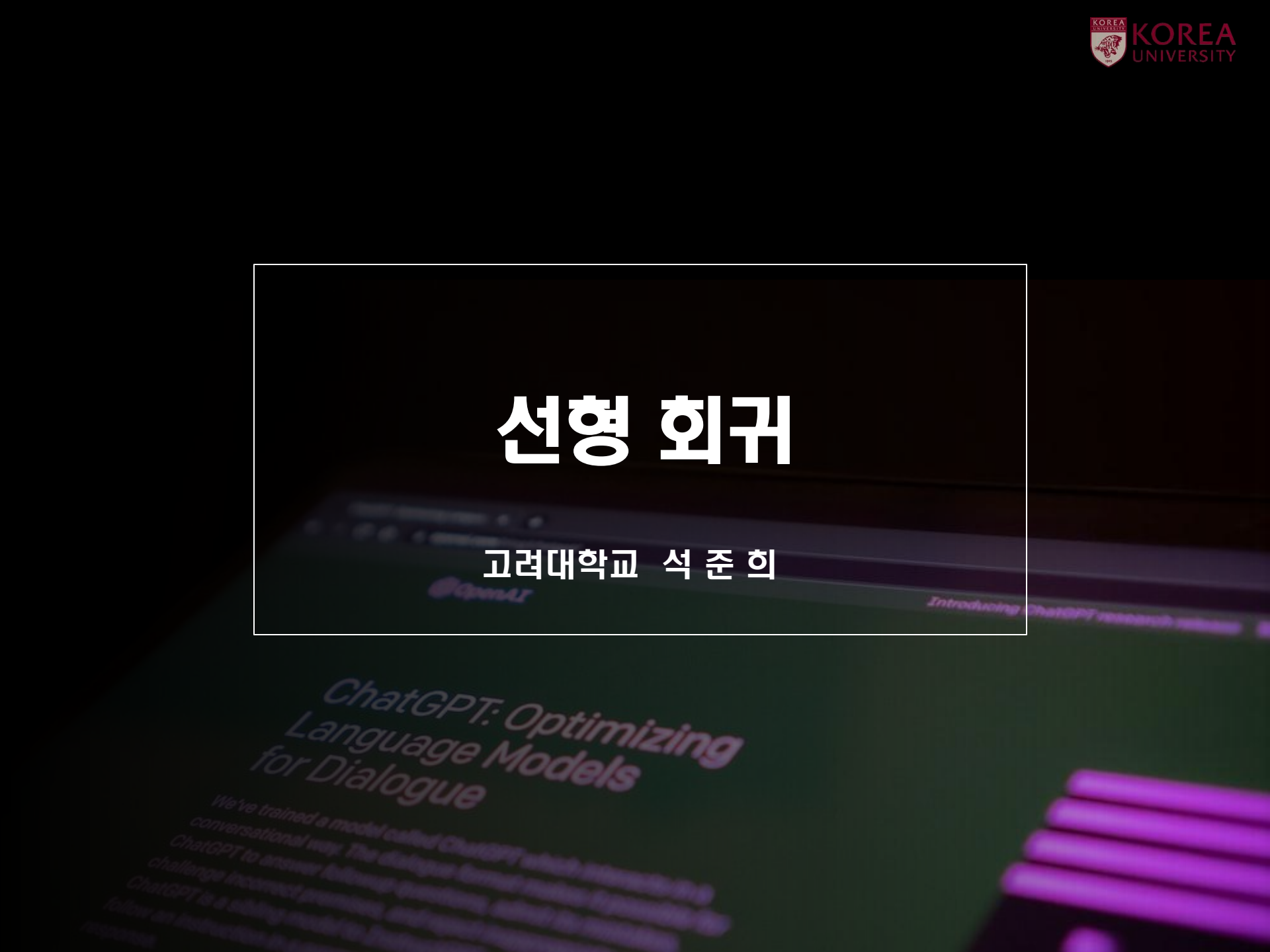
선형 회귀

고려대학교 석 준 희



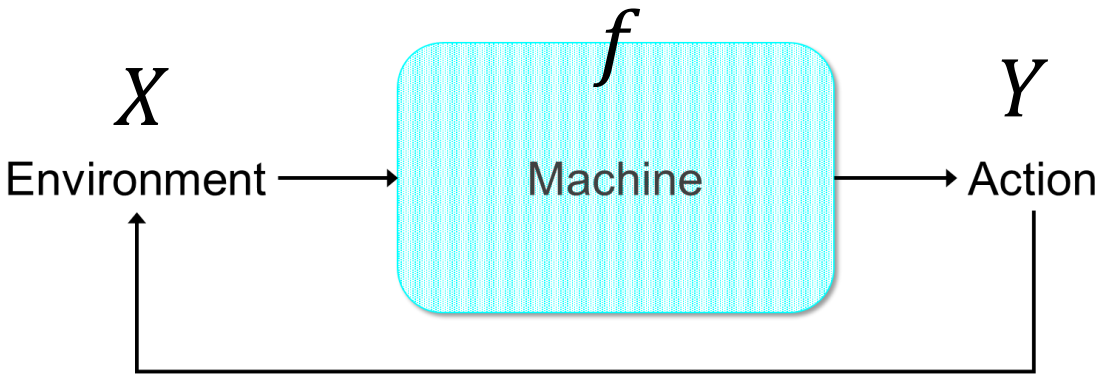
목차

선형 회귀

선형 모델

학습의 목표와 종류

- 인공지능 에이전트: 관측된 주변 환경 데이터로부터 적절한 행동에 대한 판단을 내리는 주체



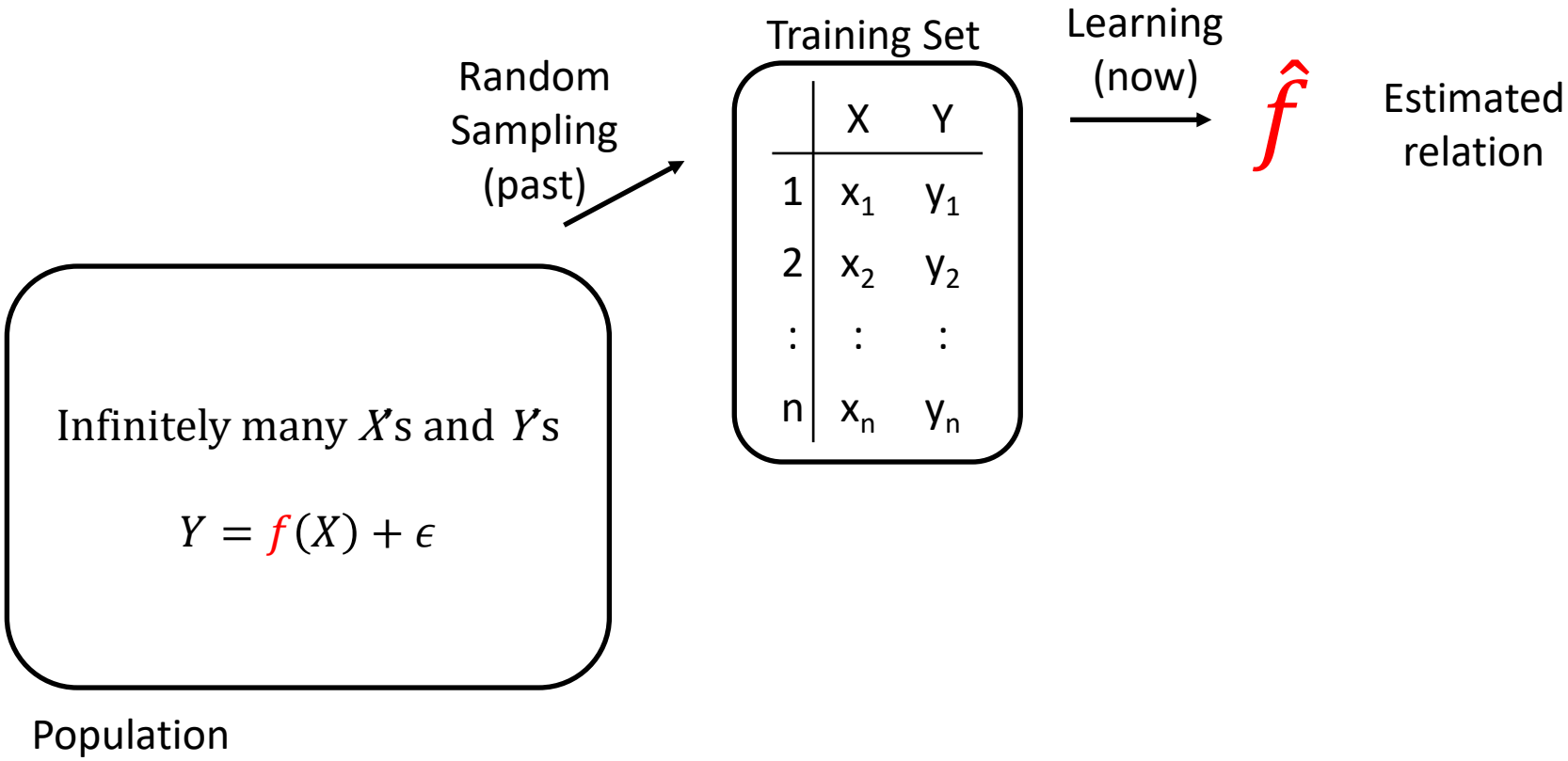
$$Y = f(X)$$

- 함수 $f()$를 주어진 관측 데이터로부터 추정하는 것이 학습의 목표
  - 지식으로부터 찾는다면 지식 기반의 인공지능

- 회귀(regression): Y가 수치형 변수일 때
  - 기대 수명, 임금, 주가, 카카오톡 메시지의 수 등
- 분류(classification): Y가 범주형 변수일 때
  - 성공/실패, 성별, 자동차의 종류, 꽃 품종 등

선형 모델 (Linear Model)

- 모집단에서의 실제 $f()$가 선형성을 갖고 있다고 가정
  - 선형 회귀 (linear regression) 모델: 회귀 문제에 적용
  - 로지스틱 회귀 (logistic regression) 모델: 분류 문제에 적용



선형 모델 vs. 최근접 이웃 기법

- 선형 모델
  - 모집단에서 선형성을 가정하는 모수적(Parametric) 접근법
  - 많은 복잡한 방법론의 기본적인 접근

- 선형 회귀/로지스틱 회귀
  - 모집단에 대한 강력한 가정을 바탕으로 한 모델
  - 모수적(Parametric) 접근: 가정으로부터 모델을 세우고 (모델)파라메터를 찾음
  - 장점: 상대적으로 적은 양의 데이터가 필요, 계산량이 적음
  - 단점: 가정이 잘 못 되면 성능의 개선이 불가능

- 최근접 이웃 기법
  - 비모수적(Non-parametric) 접근: 모집단에 대한 가정 없음
  - 유사한 X를 갖는 샘플들의 Y도 유사할 것이라는 아이디어에서 출발
  - 장점: 가정을 하지 않기 때문에 어떠한 형태의 데이터도 학습이 가능
  - 단점: 상대적으로 많은 계산량과 데이터가 필요

선형 회귀

단순 선형 회귀 (Simple Linear Regression)

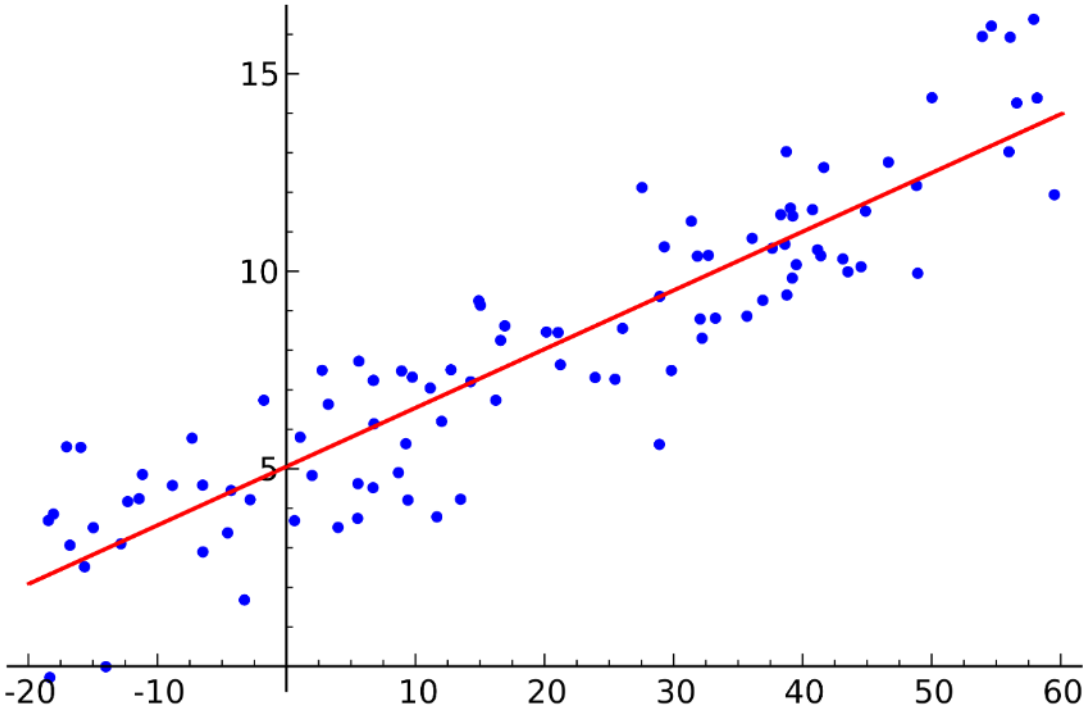
- 하나의 독립변수(X)와 하나의 출력변수(Y)에 대한 모델

$$Y \approx \beta_0 + \beta_1 X$$

β_0: 절편 (intercept, bias)
β_1: (X의) 계수 (coefficient)

- 각 데이터는 다음과 같이 모델링 됨

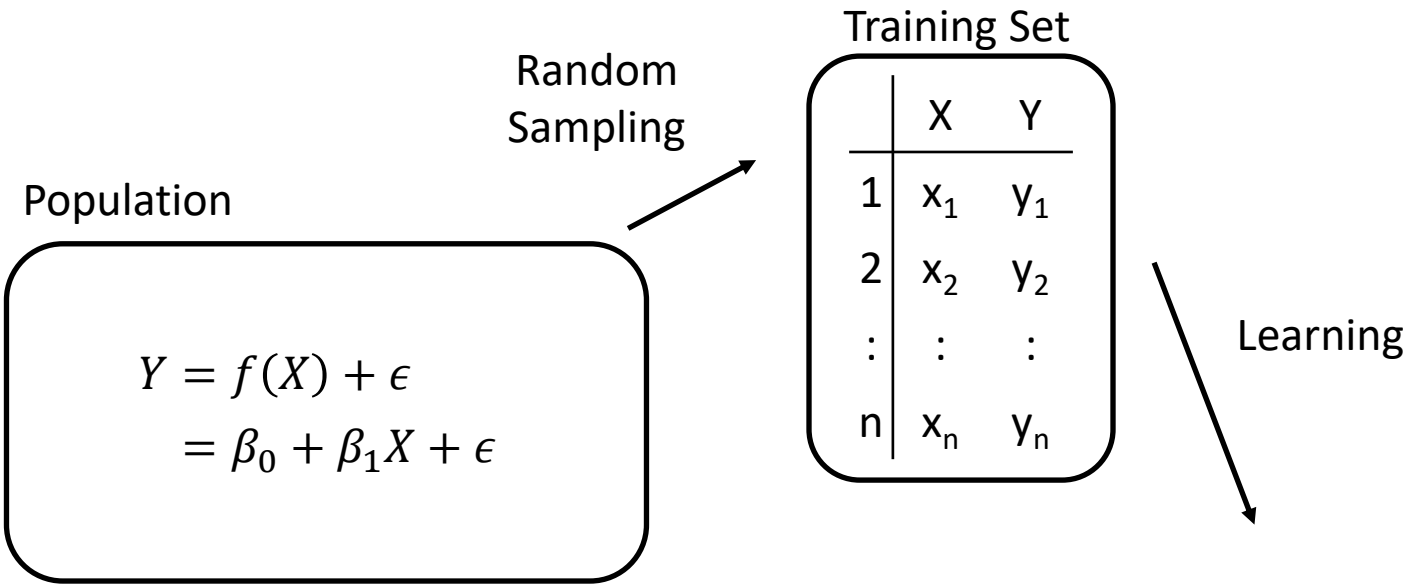
$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ for } i = 1 \cdots n.$$



https://en.wikipedia.org/wiki/Linear_regression

단순 선형 회귀 (Simple Linear Regression)

- 하나의 독립변수(X)와 하나의 출력변수(Y)에 대한 모델 $Y \approx \beta_0 + \beta_1 X$



Model Estimation $\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$

Prediction $\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

Mean Square Error $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$

모델 파라메터의 추정

- 최소제곱법(Least Square): 제곱에러 손실함수를 최소화하여 모델 파라메터를 추정

$$L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- 모델 파라메터의 추정: 손실함수를 미분하여 쉽게 추정 가능

$$\frac{\partial L(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial L(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = 0$$

⇨

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- 위의 추정값은 훈련 데이터에서의 손실을 최소화하는 것으로, 반드시 평가 데이터에서의 가장 좋은 성능을 보장하지 않음
  - 평가 데이터에서의 성능을 최대화하기 위해서는 모델 선택의 절차가 필요

일반 선형 회귀 (Multiple Linear Regression)

- 하나의 출력 변수와 여러 개의 입력 변수

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p$$

- 제곱에러(square-error) 손실함수를 최소화 하여 파라메터를 추정
  - 손실함수를 미분하여 쉽게 추정 가능

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots \hat{\beta}_p x_{pi}$$

- 일반적으로 $p+1$원 연립방정식을 풀어야함 → 행렬 형태로 간단히 표현

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$L(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{y}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$L(\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{y})/\hat{\boldsymbol{\beta}} = 0$$

⇨

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

회귀 모델의 평가

- 회귀 문제에서 추정한 모델이 얼마나 좋은지 평가하는 척도
- 평균제곱에러 (MSE: Mean square error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \qquad RMSE = \sqrt{MSE}$$

- 결정 계수(Coefficient of determination) R^2
  - Y가 얼마나 X에 의해 설명되는지에 대한 비율

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{(\text{unexplained variance})}{(\text{total variance of } Y)}$$

  - 보통은 0과 1 사이이지만, 모델이 아주 나쁜 경우 음의 값을 갖기도 함
- 이러한 척도는 훈련 데이터셋과 평가 데이터셋 양쪽 모두에서 계산 가능

회귀 모델 계산 예제

- 훈련 데이터

	X_1	X_2	Y
1	0	0	1
2	1	0	2
3	0	1	3
4	1	1	3

- 모델: $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2$

- 표본 별 모델링: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$

$$
\begin{aligned}
1 &= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + e_1 \\
2 &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + e_2 \\
3 &= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 + e_3 \\
3 &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1 + e_4
\end{aligned}
\quad \Rightarrow \quad
\begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}
$$

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}
$$

회귀 모델 계산 예제

- 손실 함수

$$L = \sum_{i=1}^{4}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$
$$= (1 - \beta_0)^2 + (2 - \beta_0 - \beta_1)^2 + (3 - \beta_0 - \beta_2)^2 + (3 - \beta_0 - \beta_1 - \beta_2)^2$$

- 손실함수 최소화를 위한 파라메터 계산

$$\frac{\partial L}{\partial \beta_0} = -2(1 - \beta_0) - 2(2 - \beta_0 - \beta_1) - 2(3 - \beta_0 - \beta_2) - 2(3 - \beta_0 - \beta_1 - \beta_2) = 0$$
$$\frac{\partial L}{\partial \beta_1} = -2(2 - \beta_0 - \beta_1) - 2(3 - \beta_0 - \beta_1 - \beta_2) = 0$$
$$\frac{\partial L}{\partial \beta_2} = -2(3 - \beta_0 - \beta_2) - 2(3 - \beta_0 - \beta_1 - \beta_2) = 0$$

$$\Rightarrow \quad \hat{\beta}_0 = 1.25, \hat{\beta}_1 = 0.5, \hat{\beta}_2 = 1.5$$

- 최종 모델

$$\hat{y} = 1.25 + 0.5x_1 + 1.5x_2$$

회귀 모델 계산 예제

- 행렬 공식을 이용한 계산

$$\begin{bmatrix}1\\2\\3\\3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\1 & 1 & 0\\1 & 0 & 1\\1 & 1 & 1\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\beta_2\end{bmatrix} + \begin{bmatrix}\epsilon_1\\\epsilon_2\\\epsilon_3\\\epsilon_4\end{bmatrix} \qquad \mathbf{y} = \mathbf{X\beta} + \mathbf{e} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix}4 & 2 & 2\\2 & 2 & 1\\2 & 1 & 2\end{bmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix}0.75 & -0.5 & -0.5\\-0.5 & 1 & 0\\-0.5 & 0 & 1\end{bmatrix}, \quad \mathbf{X}^T\mathbf{y} = \begin{bmatrix}9\\5\\6\end{bmatrix}$$

$$\Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix}1.25\\0.5\\1.5\end{bmatrix}$$

회귀 모델 계산 예제

- 모델 평가
  - 훈련 데이터 예측값 및 잔차 계산

	X_1	X_2	Y	$\hat{Y}$	e
1	0	0	1	1.25	-0.25
2	1	0	2	1.75	0.25
3	0	1	3	2.75	0.25
4	1	1	3	3.25	-0.25

학습된 모델

$$\hat{y} = 1.25 + 0.5x_1 + 1.5x_2$$

  - 훈련 데이터 평균제곱에러

$$MSE_{Train} = \frac{1}{4}(0.25^2 + (-0.25)^2 + (-0.25)^2 + 0.25^2) = 0.0625$$

  - 훈련 데이터 R^2:

$$R^2_{Train} = 1 - \frac{0.25^2 + (-0.25)^2 + (-0.25)^2 + 0.25^2}{(1-2.25)^2 + (2-2.25)^2 + (3-2.25)^2 + (3-2.25)^2} = 0.91$$

예제: 광고비와 판매량

- 상품의 판매량(Sales)을 각 매체(TV, Radio, Newspaper)의 광고비로 설명
  - $\texttt{Sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{Radio} + \beta_3 \texttt{Newspaper} + \varepsilon.$

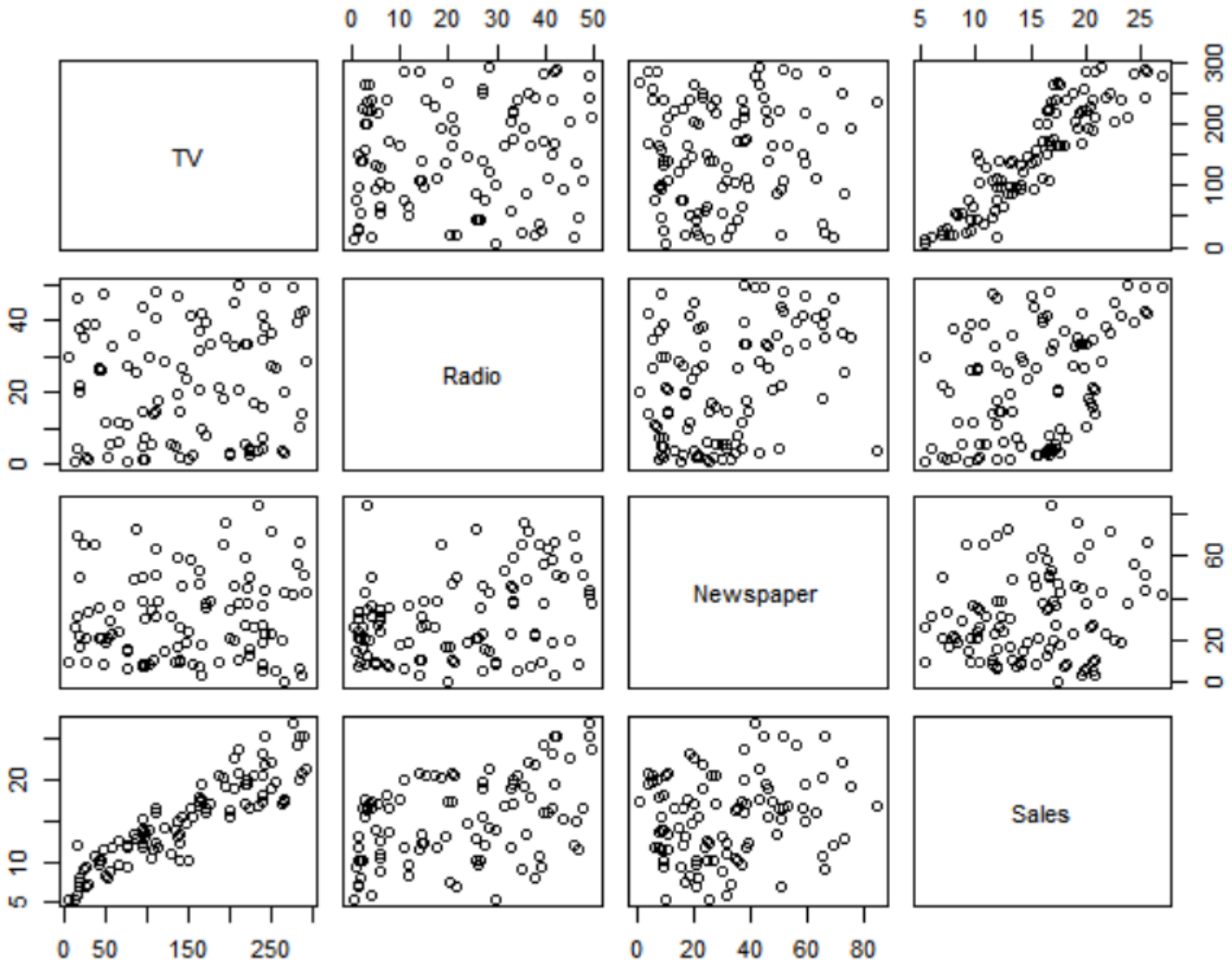
- 전체 200개의 데이터 표본 중에서...
  - 훈련 데이터셋: 임의로 선택된 100개의 표본
  - 평가 데이터셋: 나머지 100개의 표본

- 데이터 행렬

1	TV	Radio	Newspaper	Sales
2	230.1	37.8	69.2	22.1
3	44.5	39.3	45.1	10.4
4	17.2	45.9	69.3	12
5	151.5	41.3	58.5	16.5
6	180.8	10.8	58.4	17.9
7	8.7	48.9	75	7.2
8	57.5	32.8	23.5	11.8
9	120.2	19.6	11.6	13.2
10	8.6	2.1	1	4.8
11	199.8	2.6	21.2	15.6

예제: 광고비와 판매량

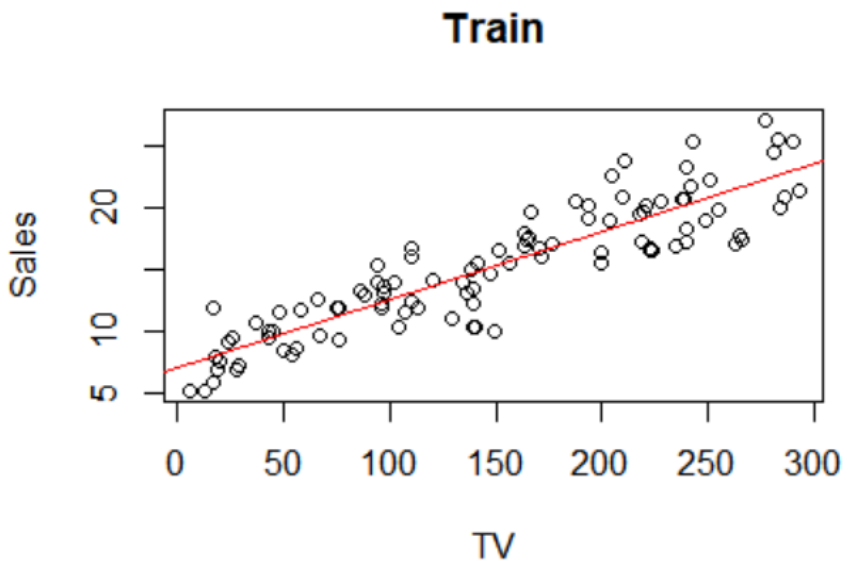
- 탐색적 데이터 분석

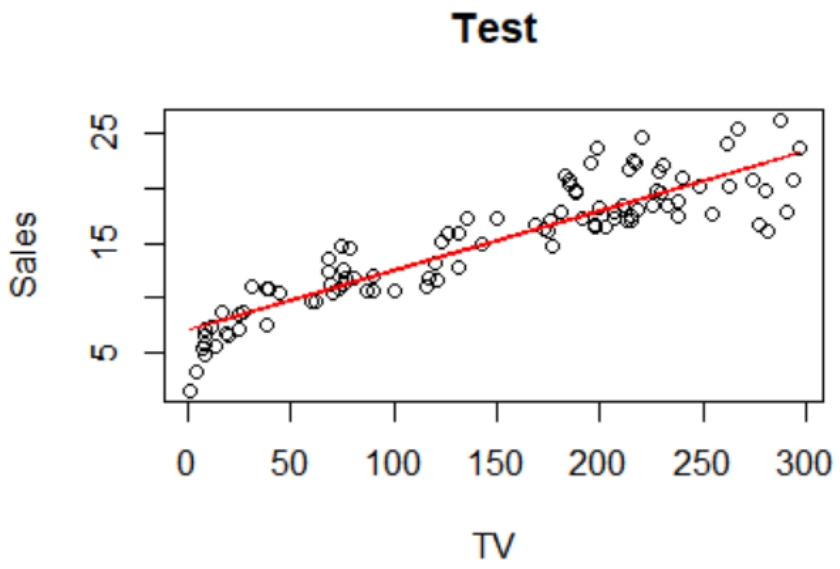


예제: 광고비와 판매량

- TV광고비만을 이용한 모델
  - $\texttt{Sales} = \beta_0 + \beta_1 \texttt{TV} + \varepsilon.$
- 결과
  - 모델: $\texttt{Sales} \sim 7.074 + 0.055\ \texttt{TV}$

	MSE	R2
Train	5.4193	0.8216
Test	5.2871	0.8017





KOREA UNIVERSITY

회귀 모델 계산 예제

- 모든 변수를 이용한 모델
  - $\texttt{Sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{Radio} + \beta_3 \texttt{Newspaper} + \varepsilon$

- 결과
  - 모델: $\texttt{Sales} \sim 4.478 + 0.054\ \texttt{TV} + 0.116\ \texttt{Radio} + 0.0002\ \texttt{Newspaper}$

	MSE	R2
Train	2.4050	0.9167
Test	3.0439	0.8858

- 단순한 모델과 복잡한 모델
  - 어느 모델이 더 복잡한가?
  - 어느 모델이 훈련 데이터셋에서 더 성능이 좋은가?
  - 어느 모델이 평가 데이터셋에서 더 성능이 좋은가? 항상 그럴 것인가?
  - 평가 데이터셋 결과를 보지 않고 평가 데이터셋에서 더 성능이 좋은 모델을 선택할 수 있을까?

선형 회귀

선형 회귀의 확장

범주형 데이터의 표현

- 입력 변수가 범주형 변수인 경우.... $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
  - X_1: 나이, 연속형 변수
  - X_2 : 성별, 범주형 변수 (남성/여성)
  - X_3 : 지역, 범주형 변수 (서울/경기/인천)

- 일반적으로 가변수(dummy variable)로 변환하여 모델링
  - 성별: $X_{2M} = 1$ (남성) or 0 (여성), 여성이 기본값(baseline)
  - 지역: $X_{3S} = 1$ (서울) or 0 (서울 외), $X_{3K} = 1$ (경기) or 0 (경기 외), 인천이 기본값

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_{2M} + \beta_3 X_{3S} + \beta_4 X_{3K}$$

남성이 여성에 비해 추가적으로 갖는 효과

- 데이터 변환 예시

	X_1	X_2	X_3
1	22	M	경기
2	28	F	인천
3	21	F	서울
4	25	M	인천

⇨

	X_1	X_{2M}	X_{3S}	X_{3K}
1	22	1	0	1
2	28	0	0	0
3	21	0	1	0
4	25	1	0	0

범주형 데이터의 표현

- 가변수 변환 vs. 인코딩(Encoding)
  - 가변수 변환은 범주형 데이터를 수치로 변환하는 인코딩의 일종
  - 일반적으로 AI에서는 one-hot encoding, label encoding 등의 활용

- 일반적으로 K개의 범주를 갖는 변수에 대해서 $K-1$ 개의 가변수가 필요
  - 성별을 나타내기 위해 $X_{2M} = 1$ (남성) or 0 (여성) 과 $X_{2F} = 0$ (남성) or 1 (여성) 을 둘 다 쓸 수는 없는가?
  - 마지막 가변수는 추가적인 정보를 제공하지 않음, e.g. 다른 가변수로부터 계산될 수 있음

- 서로 다른 방식으로 인코딩하더라도 결과는 동일함
  - 성별을 나타내기 위해 $X_{2Dummy} = 2$ (남성) or -5 (여성) 와 같은 식으로 인코딩하여도 결과는 동일

상호작용(Interaction)의 고려

- 일반적 선형 회귀 모델: 두 변수의 영향력이 독립적

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- 상호작용을 고려한 모델: 두 변수의 영향력이 비독립적

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2$$

- 상호작용 항 $X_1 X_2$ 를 추가하여 새로운 변수 X_3 처럼 취급

	X_1	X_2
1	2	-1
2	5	2
3	-3	-2

⇨

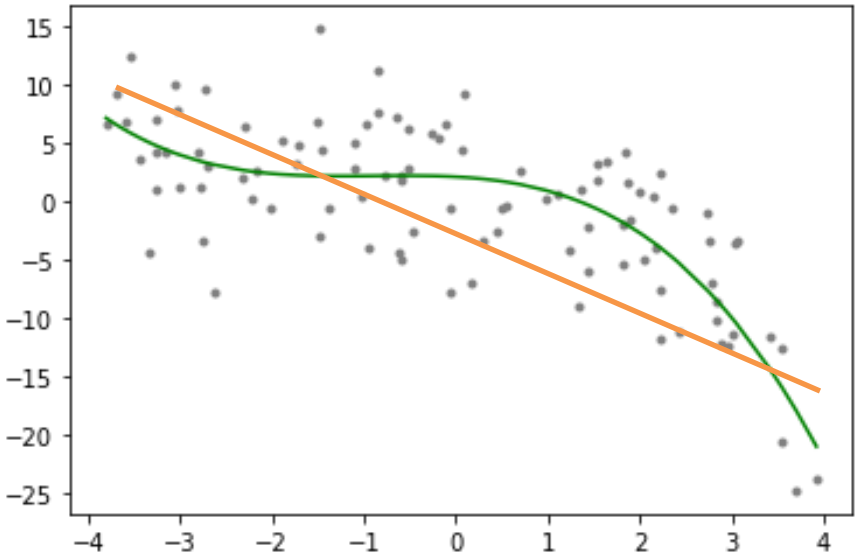
	X_1	X_2	X_3
1	2	-1	-2
2	5	2	10
3	-3	-2	6

- 상호작용을 고려한 모델 vs. 일반 모델
  - 어느 모델이 더 복잡한가?
  - 어느 모델이 훈련 데이터셋에서 더 성능이 좋은가?
  - 어느 모델이 평가 데이터셋에서 더 성능이 좋은가?

다항 회귀 (Polynomial Regression)

- X와 Y가 비선형적 관계에 있다면 선형회귀 모델이 성립하기 어려움

- 다항 회귀: 고차항을 추가하여 정합성을 높임
  - $Y \approx \beta_0 + \beta_1 X \rightarrow Y \approx \beta_0 + \beta_1 X + \beta_2 X^2$
  - 일반 선형 회귀 모델 (1차 회귀 모델) → 다항 회귀 (2차, 3차… 회귀 모델)
  - 고차항을 새로운 변수처럼 취급



	X_1
1	2
2	5
3	-3

⇨

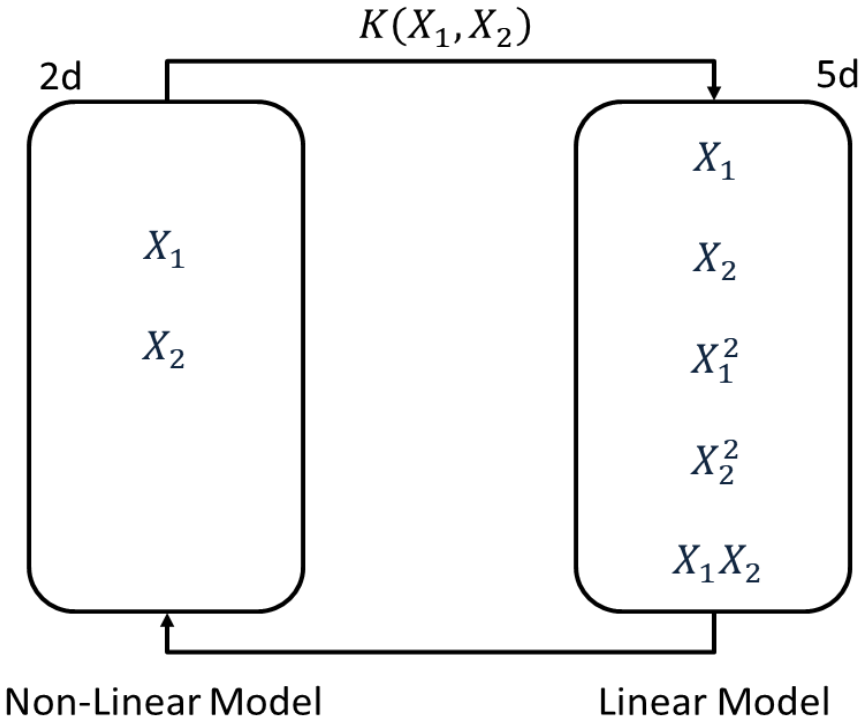
	X_1	X_2
1	2	4
2	5	25
3	-3	9

- 모델의 복잡성 vs. 고차항

비선형 모델의 해석

- 두 개의 모델에 대한 비교
  - $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$

- 원래 차원(2차원) 공간을 고차원(5차원)으로 변환하여 선형 모델을 찾음
  - 커널(Kernel) 기법: 원래 차원의 데이터를 다른 차원으로 변환하여 모델을 쉽게 찾는 기법



기계학습 모델 vs. 통계 모델

- 선형 회귀 모델은 현대의 기계학습과 전통적인 통계에서 모두 중요한 모델이지만, 이 둘은 약간 다른 관점을 갖고 있음

- 통계 모델
  - 주로 해석에 초점을 맞추고 있음
  - 적은 계산량을 필요로 하는 이론적 추정을 주로 이용
  - 상대적으로 적은 데이터 양을 가정

- 기계학습 모델
  - 주로 예측에 초점을 맞추고 있음
  - 많은 계산량을 필요로 하는 실험적 추정을 주로 이용
  - 상대적으로 많은 데이터 양을 가정

기계학습 모델 vs. 통계 모델

- 두 모델 중 어느 모델을 선택할 것인가?
  - $f_1()$: $\texttt{sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{radio} + \beta_3 \texttt{newspaper} + \varepsilon$
  - $f_2()$: $\texttt{sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{radio} + \varepsilon$

- 통계 모델
  - 모델의 파라메터가 0인지 아닌지 (이론적) 통계적 검정을 이용해 확인

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

  - 모델의 실제 에러를 이론적으로 추정: $\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

- 기계학습 모델
  - 모델의 예측 성능을 다른 데이터를 이용해 확인 (평가 데이터)

KOREA UNIVERSITY

감사합니다