Supervised learning I.

Linear Regression : 회귀의 기본 모델.
선형 회귀 모델의 파라미터인 계수에 대해서
통계적 테스트를 진행할 수 있음.
유용한 변수 선택에 대해 알아볼 수 있음.
비선형적인 데이터를 어떻게 적합할지 알수있음.

β_1을 유의한가?
β_1은 X의 계수로 0인 경우 Y와 무관해짐.
이를 T-test로 확인할 수 있음.
$H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$, t값을 구하고,
T분포 내에서 t값을 통해 p-value를 찾을수 있음

회귀는 데이터 포인트를 가장 잘 설명하는 회귀선을
찾자 하는 것임. 모집단에는 Y를 제대로 설명하는
회귀계수와 편향이 없으나, 이를 확인할 수는 없음.
그래, 표본의 MSE를 통해 회귀계수를 찾아야 함.
RSS는 잔차제곱합으로, MSE에서 n만을 제거한것.
$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2 = \sum_{i=1}^{n} e_i^2$
RSS가 2차 함수이므로 이를 미분하여 각 계수에 대한 최적의 값을 찾아낼 수 있음.
이렇게 만들어진 회귀 계수는 표본으로 부터 얻은 것이기 때문에 모집단의 회귀계수의 추정값이됨
추정된 회귀계수와 모집단의 회귀계수가 얼마나 동일한지는 통계적 문제임.

모델과 데이터의 적합도를 측정하는 지표
$RSE = \sqrt{\frac{RSS}{n-2}}$ (MSE의 좋은 추정치임)
$R^2 = 1 - \frac{RSS}{TSS}$
→ TSS는 총제곱합, RSS는 오차 함계의 제곱
R^2이 1에 가까울 수록 좋음.

다변량 선형회귀
$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
→ 여기에서 RSS를 최소화하는 $\hat{\beta}$를 찾음.
$y = X\beta + e$
→ $y = \begin{bmatrix} y_1 \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_p \end{bmatrix}$, $e = \begin{bmatrix} e_1 \\ e_n \end{bmatrix}$
$RSS = (y-\hat{y})^T (y-\hat{y})$
$\hat{\beta} = (X^T X)^{-1} X^T y$
$\hat{y} = X (X^T X)^{-1} X^T y$
$RSE = \sqrt{\frac{RSS}{n-p-1}}$
SE matrix를 통해 각 β의 SE를 얻을 수 있음.
β_0는 $\beta_0 \pm 1.96 SE(\hat{\beta}_0)$ 사이에 95% 있음
이후 T-test를 통해 β_0가 0인지 확인함.

표본으로 찾은 $\hat{\beta}$과 모집단의 β
μ를 기준으로 신뢰구간을 찾을 때는 아래와 같음.
$p = Pr[\hat{\mu} - k\,SE(\hat{\mu}) < \mu < \hat{\mu} + k\,SE(\hat{\mu})]$
T-test를 통해서 실제 μ와 $\hat{\mu}$인지를 확인함.
T-test는 $(\hat{\mu} - \mu_0)/SE(\hat{\mu})$ 값을 관찰할 확률을
T분포에서 확인함으로서 p-value를 구함.
이런 신뢰구간을 통해서 모집단의 회귀계수와
$\hat{\beta} \pm k\,SE(\hat{\beta})$ 범위 내에 몇 %로 존재한다고
결론 내릴 수 있음

변수 선택의 문제.
변수가 많을 수록 유연성이 높아짐.
훈련세트에서 RSS를 줄일수 있다. 과대적합되기
쉬운 문제점이 있음.
어떤 변수가 유용한지 파악하고 이것들만
사용하는 것이 정확함.