Regression

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

$$E[(Y - \hat{f}(X))^2 | X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}$$, when $\epsilon = Y - f(X)$

Curse of dimensionality: a neighbor in high-dimension need no longer be local(sparse)-"neighborhood" becomes less meaningful (getting apart)

Bayes classifier has smallest error in population using true Pk(x), Bayes error: lowest test error, but true probability is unknown, KNN: k-hyperparameter

Linear function with error (Y: dependent, output, response, target / X: independent, input, predictor, feature)

$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e$ / **β0: intercept, βi: slope, e: error**

Bias-Variance Trade-off

$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \mathrm{Var}(\epsilon)$ / $\mathrm{Bias}\left(\hat{f}(x_0)\right) = E\left(\hat{f}(x_0)\right) - f(x_0)$ / $\mathrm{Var}[\hat{f}(x)] = \mathrm{E}\left[\left(\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\right)^2\right]$

더 많은 데이터를 포착하려 더 많이 움직인다면(Flexible) bias↓, variance↑

Degree of Freedom: the number of independent pieces of information

sample variance will be what is called an Unbiased estimator of the population variance σ^2

Least Squares estimator

$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$ / $\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$ / minimizing SS condition is $B = \frac{S_{xY}}{S_{xx}}$, $A = \overline{Y} - B\overline{x}$ / $\sum_{i=1}^{n} \frac{(Y_i - E[Y_i])^2}{\mathrm{Var}(Y_i)} = \sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2} \sim \chi^2_n$

Sample means **Variance**

(normal distribution iid random variables(Unbiased estimator))

$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ $\quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$ or $\mathrm{Var}(B) = \frac{\sigma^2}{S_{xx}}$, $\quad \mathrm{Var}(A) = \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\bar{x}^2$

Example for H0(null hypothesis H0 : β1=0)

$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$ / H0: there's no relationship between X and Y

- Significance level γ test of H_0

$$\text{reject} \quad H_0 \quad \text{if} \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|B| > t_{\gamma/2, n-2}$$
$$\text{accept} \quad H_0 \quad \text{otherwise}$$

v value: $\sqrt{(n-2)S_{xx}/SS_R}|B|$ = point

$p\text{-value} = P\{|T_{n-2}| > v\}$ = 면적

$= 2P\{T_{n-2} > v\}$

$P\{t_{30} > 2.04\} = 0.025, \; P\{t_{30} > 2.75\} = 0.005$ / 95%(2.04): $\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \; \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$

Confidence interval estimator for β of SSR

$P\left\{B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2} < \beta < B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right\} = 1 - a$ / **σ(error에 대한 분포)를 모르기 때문에 ~t distribution,** 안다면 **~N**

H0:α1=0 & Confidence interval estimator for α of SSR

$\sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_i x_i^2}}(A - \alpha) \sim t_{n-2}$ / $A \pm t_{\alpha/2,n-2}\sqrt{\frac{SS_R \sum_i x_i^2}{n(n-2)S_{xx}}}$ = 1-a

Inferences About	Use the Distributional Result
β	$\sqrt{\frac{(n-2)S_{xx}}{SS_r}}(B - \beta) \sim t_{n-2}$
α	$\sqrt{\frac{n(n-2)S_{xx}}{\sum_i x_i^2 SS_R}}(A - \alpha) \sim t_{n-2}$

Residual Standard Error(RSE) / (SSR: input 빼고 error만 고려하자)

$\mathrm{RSE} = \sqrt{\frac{1}{\underbrace{n-2}_{\text{unbiased estimator}}}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ / $\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$ (RSS감소가 p에 비해 미미하면 RSE는 more variable higher)**or**

$E\left[\frac{SS_R}{n-2}\right] = \sigma^2 \quad \frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$

R-squared(R^2) and sample correlation coefficient(only for a single): 설명가능한 정도/ 0~1 클수록 better fit of the model to the data / more variable, always increase / 얼마나 학습을 잘했는가?

- Coefficient of determination

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}}$$

$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$ **or** $SS_R = \sum_{i=1}^{n}(Y_i - A - Bx_i)^2 \quad S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ sample correlation coefficient : $\boxed{r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\underbrace{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}_{\text{Correlation}}} = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}}$ $|r| = \sqrt{R^2}$ (-1~1)

multiple linear regression: x 전부를 고려하며 나머지는 임의의 값 **vs simple linear regression:** x: 1개만 고려하고 나머지는 0

ideal scenario: the predictors are uncorrelated(i.i.d.) but claims of causality

	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

coefficient: βi

std.error: standard deviation of estimated βi

t-statistic: normalized random variables for H0, βi assuming their zero means / normal dist / only can use sample mean&variance, cause they are unknown / 개별 변수의 영향력(t-statistic, q=1일때의 f dist), 전체 변수의 영향력(f-statistic for a large number of features even though p-values are small) / 나머지를 모두 0으로 만든게 아닌 특정값을 넣고 계산(**single**과의 차이) z-statistic(binomial일때) vs t-statistic(random)

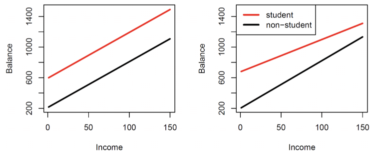
p-value: probability of t-statistics / H0의 accept reject 결정 / 클 수록 non-significant 작을수록 significant

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

즉, newspaper는 단독적으로는 의미가 매우 적지만, radio에 큰 영향을 받아 sales에 영향을 미친다 (**correlation**)

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \boxed{\beta_3 \times \texttt{income}_i} & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$



Is at Least one Predictor useful?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

$n-(p+1)$ (n: sample 수, p: coefficient 수)

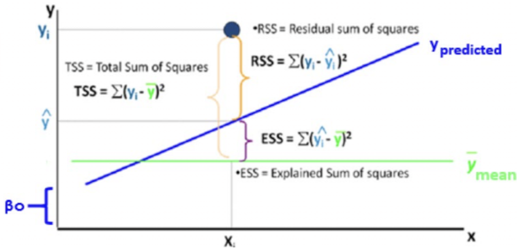
Explained Sum of Squares (ESS) = $\sum(\hat{y}_i - \bar{y})^2$ -> (TSS-RSS)

$$\boxed{E\{\text{RSS}/(n-p-1)\} = \sigma^2}$$

H_0 true: $\boxed{E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2}$

H_a true: $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$

ESS



Deciding on important variables?

Forward selection: lowest RSS, stopping rule, greedy
Backward selection: largest p-value is removed

How well does model fit data?

R^2 or RSE

What Response value and how accurate?

confidence interval: average response / prediction interval: individual response
confidence interval은 평균을 이용하므로 sample의 수가 더 많다 이때, Random error의 효과가 서로 상쇄된다.
즉, confidence interval이 prediction interval보다 항상 좁다 (=더 정확한 값을 얻을 수 있다는 이야기)

Interaction(Synergy)

한 축에 치우쳐서 평가하면 overestimate됨 / 두 변수의 곱으로 표현되는 변수 추가 ex) (radio * TV) or R^2: (96.8-89.7)/(100-89.7)=69% of variability in sales

hierarchy: hard to interpret in a model without main effects 즉, 매우 작은 p-value를 갖더라도 기존 variable을 유지해야한다.

correlation of error terms: ρ가 커질수록 smooth / heteroscedasticity: 한쪽의 값이 커지면 variance가 커진다(키-몸무게) -> log 적용으로 해결

Outlier: x는 잘 있는데 y가 튐 (Affecting RSE, confidence interval and p-values) (너 말고도 많아)

High Leverage point: y는 잘 있는데 x가 튐 / 개별 predictor만으로는 판단X, 전체 predictor를 고려(얘 없음 안돼)

Logistic Regression(Classification): linear regression은 0보다 작거나 1보다 큰 값을 뱉음, estimate를 잘 못함 / nearest-neighbor averaging(if 0 prob)

(Logit)
$0 < \text{odd} < \infty$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \;/\; \boxed{\log\left(\boxed{\frac{p(X)}{1 - p(X)}}\right) = \beta_0 + \beta_1 X}$$

- 사실상 probability (0~1이므로) -> get qualitative response Y(classifying)

Maximum likelihood

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

probability

: coefficient estimate, maximization

Discriminant Analysis(LDA, QDA, Naive bayes): bayes theorem기반 / normal (gaussian) distribution for each class

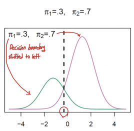
: more than 2 classes / small n / classes are well separated / two-classes일 경우 LDA=Logistic regression

posterior, Likelihood, Prior

$$\boxed{\Pr(Y = k | X = x) = \frac{\Pr(X = x | Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}} \rightarrow \boxed{\Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}}$$

posterior, Prior, Likelihood

(DA)- **maximize posterior**



Likelihood / Gaussian

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2}$$

(이때, σ=σk) 즉,

Posterior, Prior

$$\max\uparrow p_k(x) = \frac{\pi_k \boxed{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

↑max (K를 선택: LDA 목적)

Largest discriminant score

$$\max\uparrow \boxed{\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)}$$

- which is linear function / 만약 k=2, π1=π2=0.5일 때의 bayes decision boundary $x = \frac{\mu_1 + \mu_2}{2}$

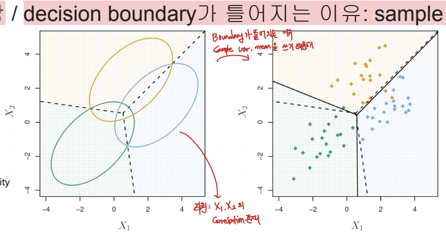
parameter: 알 수 없으므로, estimate값을 써야함 / decision boundary가 틀어지는 이유: sample variance, mean을 쓰기 때문에

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^{K} \boxed{\frac{n_k - 1}{n - K}} \cdot \hat{\sigma}_k^2$$

Unbiased estimator
weights

Ellipses: 95% of probability



softmax

$$\widehat{\Pr}(Y = k | X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

/ False negative: true인데 negative로 판단 / Threshold를 수정해서 overfitting을 막을 수 있음

LDA vs QDA vs Naive Bayes

same covariance matrix ∑ (LDA - 1차식) vs different own covariance matrix ∑ (QDA - 2차식, overfitting in linear, flexible) vs diagonal matrix ∑ (all independent)