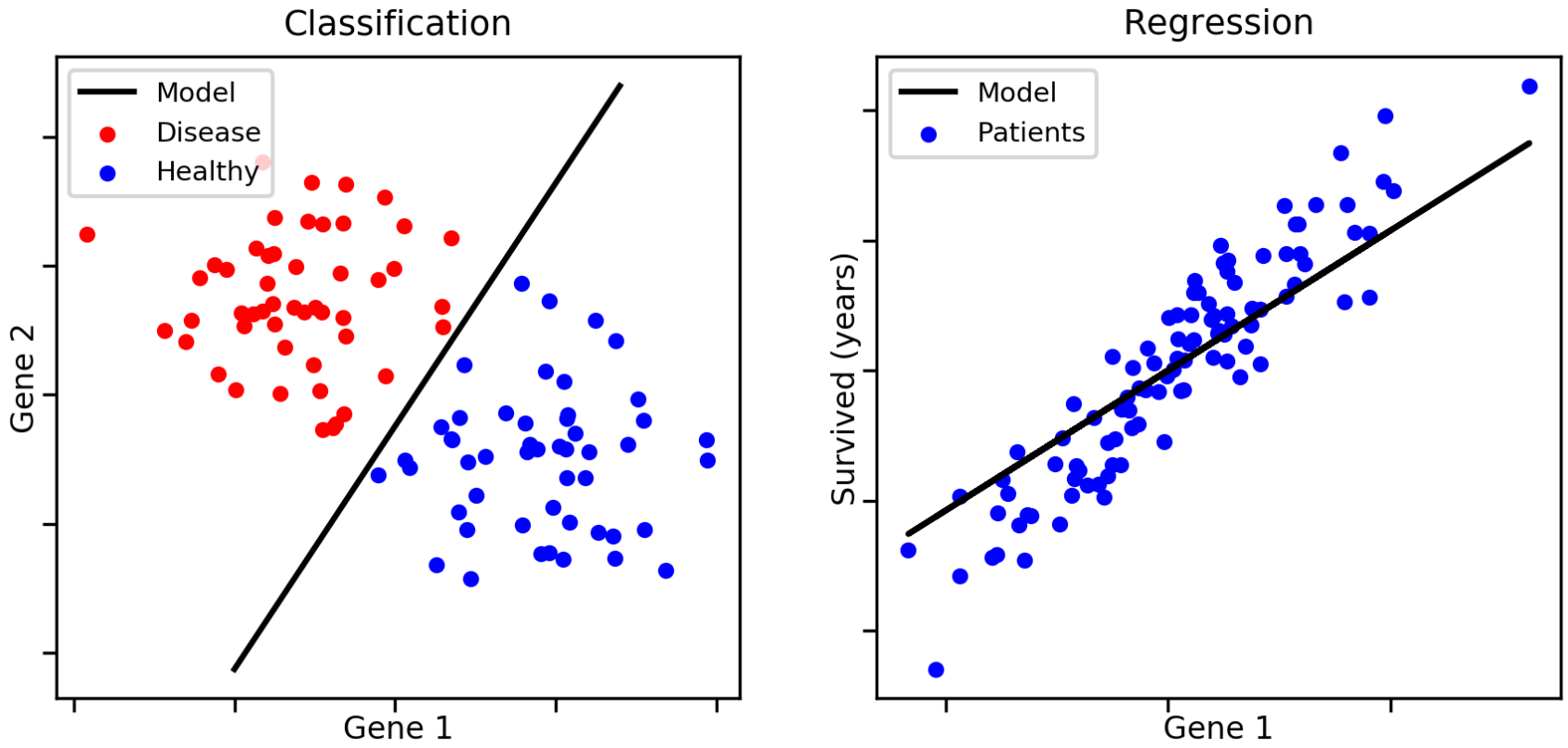
Linear Regression for Quant

장 연 식

통계 예측 모형

데이터(정답)가 주어진 문제를 알고리즘으로 푸는 통계 모델링 방식을 Supervised Learning이라고 한다.



Classification (분류) : Discrete한 값으로 주어지는 **Qualitative** target value를 예측 하는 알고리즘

Regression (회귀) : Continuous한 값으로 주어지는 **Quantitative** target value를 예측 하는 알고리즘

Linear regression (선형회귀)

독립변수와 설명변수 간의 선형 관계를 파악한다.

- 선형회귀는 대표적인 수치예측 방법이다.
- 한 개 이상의 **독립변수** (independent variable) 존재 : $X_1, X_2, \ldots, X_p$
  (= explanatory variable = exogenous variable = factor = feature = regressor)
- 한 개의 **종속변수** (dependent variable) 존재 : Y
  (= response variable = endogenous variable = predicted = outcome = regressand)
- 독립변수와 종속 변수 간의 관계를 선형 관계로 표현:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_p + \varepsilon$$
$$= X\beta + \varepsilon \text{ (matrix notation)}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

n x 1 | n x (k+1) | (k+1) x 1 | n x 1

- Explained : $\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
- Unexplained : ε *(error)*

- 오차항 ε 가 평균 = 0, 표준편차 = σ (상수)인 Random noise로 주어지게 된다면 해당 선형회귀 모델은 올바르다고 할 수 있다.

Ordinary Least Square (최소자승법; OLS)

오차제곱합을 최소화 시키는 estimator를 계산한다.

목적 함수를 최소화 시키는 방법

$\mathbf{X}\beta = \mathbf{y}$ 로 fitting 할 수 있다고 가정하면,

잔차 제곱합을 목적 함수로 최소화 시키는 **beta**를 찾는다.

$$\hat{\boldsymbol{\beta}} = \arg\min S(\boldsymbol{\beta})$$

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

$$\frac{d\boldsymbol{S}}{d\boldsymbol{\beta}} = \frac{d}{d\boldsymbol{\beta}}[(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})]$$

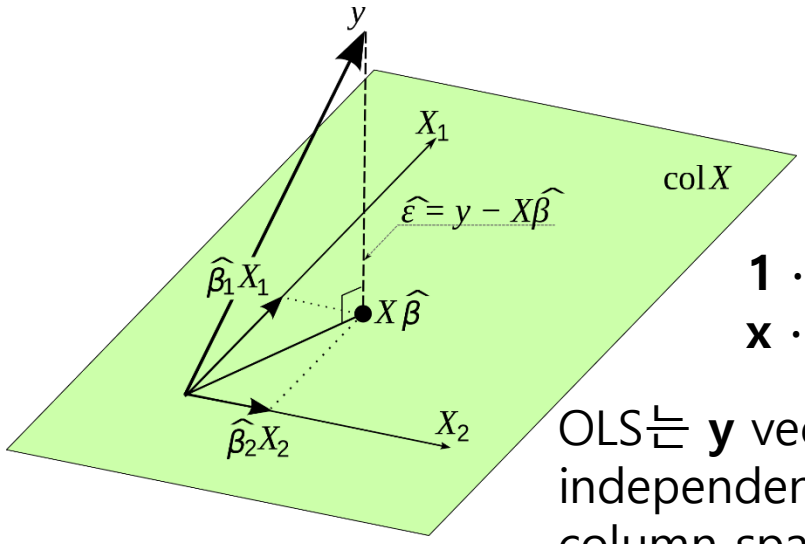
$$= \frac{d}{d\boldsymbol{\beta}}[-2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X\beta})] = 0$$

X vector들이 linearly independent라고 가정하면,

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

"Normal equation"

기하학적 해석



$\mathbf{1} \cdot \hat{\boldsymbol{\varepsilon}} = \sum \hat{\varepsilon}_i = 0$

$\mathbf{x} \cdot \hat{\boldsymbol{\varepsilon}} = \sum \mathrm{x}_i \hat{\varepsilon}_i = 0$

OLS는 **y** vector를 linearly independent **X** vector의 column space로 **orthogonal projection** 한 것과 동일한 estimation이다!

$$(\mathbf{y} - X\hat{\boldsymbol{\beta}})^{\mathrm{T}} X = 0.$$

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

OLS 기하학적 해석의 의미

OLS는 특정 독립 변수의 영향을 제회한 종속 변수를 나머지 독립 변수에 대해 projection한 결과이다.

$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \boldsymbol{\varepsilon}$, where $\mathbf{X} \equiv \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} = (\boldsymbol{X_1} \quad \boldsymbol{X_2})$ (matrix block으로 표현하자)

matrix block의 성질 중 $\boldsymbol{X}^T = \begin{pmatrix} \boldsymbol{X_1^T} \\ \boldsymbol{X_2^T} \end{pmatrix}$ 을 이용하여 normal equation을 표현하면,

$$\begin{pmatrix} \boldsymbol{X_1^T X_1} & \boldsymbol{X_1^T X_2} \\ \boldsymbol{X_2^T X_1} & \boldsymbol{X_2^T X_2} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X_1^T y} \\ \boldsymbol{X_1^T y} \end{pmatrix} \Rightarrow \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X_1^T X_1}\right)^{-1} \boldsymbol{X_1}(\boldsymbol{y} - \boldsymbol{X_2}\hat{\boldsymbol{\beta}}_2) \\ \left(\boldsymbol{X_2^T X_2}\right)^{-1} \boldsymbol{X_2}(\boldsymbol{y} - \boldsymbol{X_1}\hat{\boldsymbol{\beta}}_1) \end{pmatrix}$$

즉, $\hat{\boldsymbol{\beta}}_{\mathbf{1(2)}}$는 $\boldsymbol{X}_{\mathbf{2(1)}}$의 **영향을 제외한 y**를 $\boldsymbol{X}_{\mathbf{1(2)}}$의 column space에 projection한 결과이다.
X vector들이 orthogonal 하다면, $\hat{\boldsymbol{\beta}}_{\mathbf{1(2)}}$는 $\boldsymbol{y}$를 $\boldsymbol{X}_{\mathbf{1(2)}}$ 하나에만 projection한 것과 같다.

⇒ Data cleaning 작업 중, 특정 변수 (ex. $\boldsymbol{X_1}$)의 영향을 제거하기 위해서 ($\boldsymbol{X_1}$, $\mathbf{y}$)로 첫 번째 regression을 돌린 후 남은 값을 종속 변수로 하여 나머지 독립 변수에 대해 regression하면 된다. (2-Stage-Least-Squares; **2SLS**)

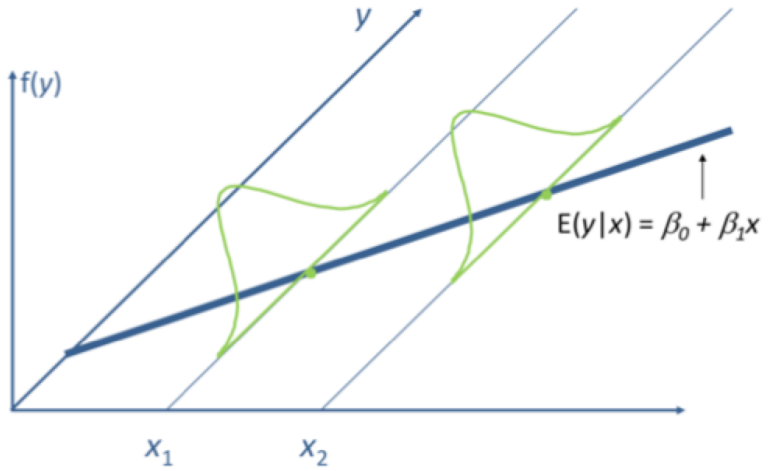
- (ex1) 포트폴리오의 성과 평가 중 특정 팩터의 영향을 제거한 성과를 측정하고 싶다면? 특정 팩터에 대해서만 regression한 후 남은 값으로 성과 평가를 한다.
- (ex2) 분기실적 발표시 주가의 earning shock가 있다면? Estimation의 정확성이 크게 떨어질 것이므로 earning shock에 해당하는 변수에 대해 regression한 후 남은 데이터로 분석을 한다.

OLS as Best Linear Unbiased Estimator

Gauss-Markov 가정 하에서 OLS estimator는 BLUE다.

Gauss-Markov 가정

1. 모집단의 독립변수와 종속 변수 간의 관계는 선형
2. Random i.i.d. sampling (cross sectional data에서 해당)
3. Stationary and ergodic stochastic process (time series data에 해당)
4. 독립변수간의 선형 독립 i.e., rank($\mathbf{X}$) = k + 1
5. Strict exogeneity i.e., $E(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0$
6. 동분산: $\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2$ 및 자기상관 없음: $\text{Cov}(\epsilon_i, \varepsilon_j) = 0 \ \forall \ i, j$

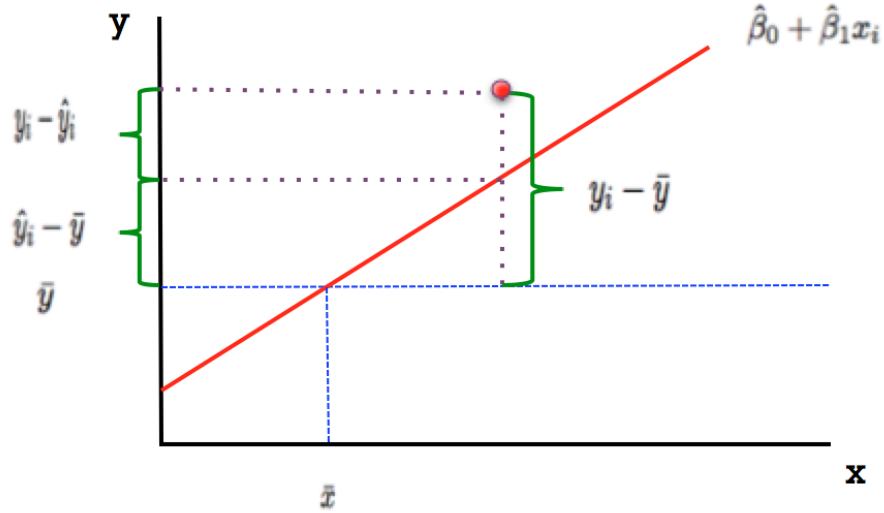


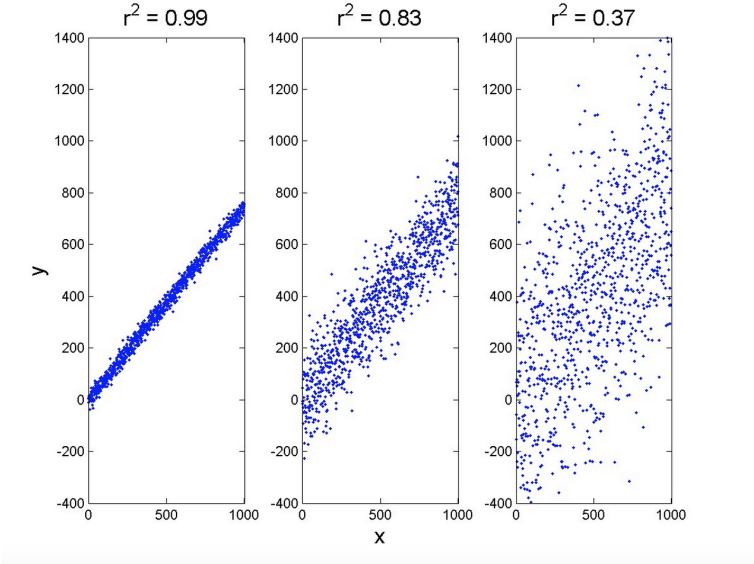
이 때, **OLS 추정량**은 가장 분산이 작은 (**B**est) 선형 (**L**inear) 불편 (**U**nbiased) 추정량 (**E**stimator)이 된다!

- OLS의 불편성 : $\boldsymbol{E(\hat{\beta})} = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ $= E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X\beta} + \boldsymbol{\varepsilon})] = \boldsymbol{\beta}$
- OLS의 분산 : $\boldsymbol{Var(\hat{\beta})} = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{\varepsilon\varepsilon}^T)\ (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{\sigma^2 I(X^TX)^{-1}}$, $Var(\hat{\boldsymbol{\beta}}_j) = \dfrac{\boldsymbol{\sigma}^2}{\boldsymbol{SST_j}\left(1-R_j^2\right)}$
- 모분산의 불편추정량 : $\hat{\boldsymbol{\sigma}}^2 = \dfrac{\boldsymbol{e}^T\boldsymbol{e}}{n-k-1}$ (e는 잔차 벡터)

OLS의 적합도

모집단의 분산 중 OLS로 설명되는 선형관계가 차지하는 비중을 R-square라고 한다.





$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum (y_i - \hat{y}_i)^2$$

$$SST = SSE + SSR$$

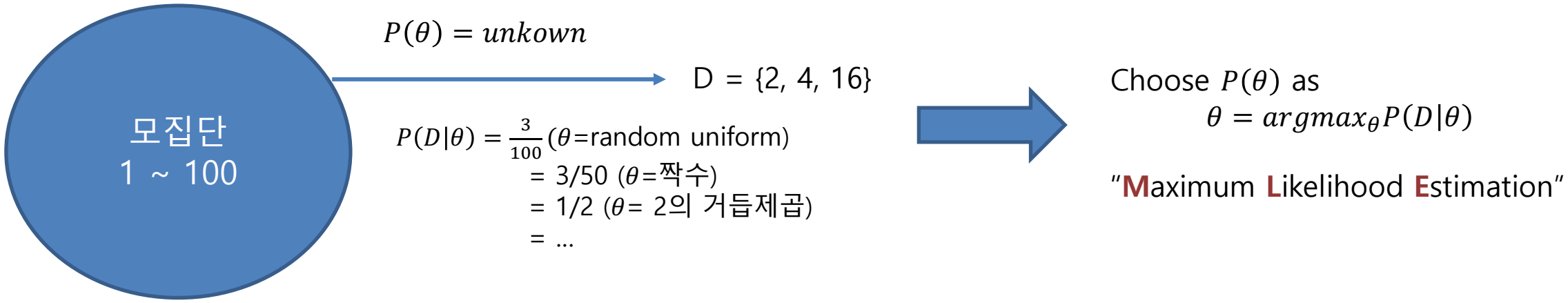
$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2}$$

$* \dot{} \; \underbrace{Var(Y)}_{\text{n * SST}} = \underbrace{E_X(Var(Y|X)}_{n * \sigma_\varepsilon^2} + \underbrace{Var_X(E(Y|X))}_{\text{n * SSE}}$ 임을 이용함

모집단 내에서 X 변화에 의해 설명되는 y 변화의 정도

Maximum Likelihood Estimation (최우추정법)

데이터의 분포가 정해져 있을 때 최고의 추정방법.



- 표본 데이터 D가 주어졌을 때, D를 조건으로 하는 사후 확률 $P_{post}(\theta)$는 베이즈 정리에 의해 $P_{post}(\theta) = P(\theta|D) = \frac{P(D|\theta)P_{prior}(\theta)}{P(D)}$ 이므로, $P_{post}(\theta) \propto P(D|\theta)$가 성립한다. 따라서 사후 확률을 최대화 시키는 확률 분포로 최적화 시키고자 한다면 **likelihood $P(D|\theta)$를 최대화 시키는 θ를 찾으면 된다**.
- D가 정해진 확률분포 $P(\theta)$로부터 random sampling 된 것이라면, MLE는 다른 어떤 추정보다도 우수한 최고의 추정 방법이다.

OLS와 정규분포

데이터가 정규 분포를 따르면 OLS는 MLE 추정량과 동일하며 모든 추정중 최고이다.

주어진 표본 D에 대해 OLS 추정을 했고, 오차항이 정규 분포 ~ $N(0,\sigma^2)$를 따른다고 하면,
추정량 $\hat{\beta}$ 또한 정규분포를 따르게 된다. 정규 분포의 선형 결합은 정규 분포이므로 따라서 **모집단 데이터 분포 또한 정규 분포**이다. 따라서 **MLE추정이 최고이다**.

정규 분포에 대한 Likelihood를 계산해 보면, $p(y|\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta))$ 이고,

Log-likelihood를 최대화 시키는 estimator를 계산하면 된다.

$$max\ lnL = -\frac{n}{2}ln2\pi - \frac{n}{2}ln\sigma^2 - \frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)$$

$$\frac{\partial}{\partial\hat{\beta}}lnL = -\frac{1}{\hat{\sigma^2}}(-X'y + X'X\hat{\beta}) = \frac{1}{\hat{\sigma^2}}(X'y - X'X\hat{\beta}) = 0$$

$$\frac{\partial}{\partial\hat{\sigma^2}}lnL = -\frac{n}{2\hat{\sigma^2}} + \frac{1}{2\hat{\sigma^4}}(y-X\beta)'(y-X\beta) = 0$$

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y$$

$$\hat{\sigma^2}_{ML} = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n} = \frac{e'e}{n}$$

"데이터가 **정규 분포**를 따르면 MLE 추정량과 OLS 추정량은 같으며 따라서 OLS는 (비선형 포함) 모든 추정중 최고의 방법!

- 데이터가 분포함수로부터 나오는 랜덤변수가 아닐때 (MLE가 최선이 아니게 됨), 강한 패턴 (비선형성)을 가지고 있을때는(Analytic한 최적값 계산이 어려움) **머신러닝** 방법론이 적절함.

OLS의 분포추론

데이터가 정규 분포를 따르면 OLS 추정량으로 일반적인 t, F 가설검정을 할 수 있다.

Gauss-Markov 가정이 성립하고 데이터(오차)의 정규성이 있는 경우 앞선 논리(정규 분포의 선형성)로 부터 추정량의 분포 또한 정규 분포이다.

$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$, where $Var(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj} = \frac{\sigma^2}{\sum (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}$

따라서 이를 이용하면 t-검정이 가능하다.

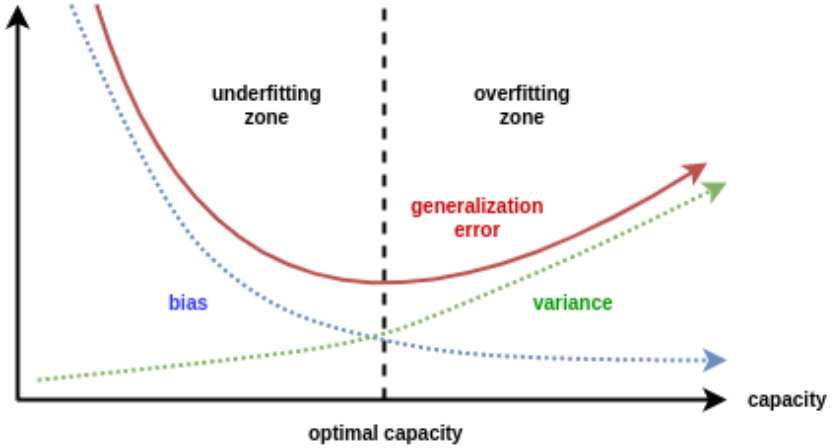
$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t(n-k-1), \text{ where } se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\sum (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}}, \quad \hat{\sigma} = \sqrt{\frac{e'e}{n-k-1}}$$

결합 가설 검정을 위한 F-statistics 도 만들 수 있다.

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\frac{e'e}{n-k-1}} = \frac{\hat{\beta}_j{}^2 / (X'X)^{-1}_{jj}}{e'e/n-k-1} \sim F(1, n-k-1)$$

OLS의 과소적합 (underfitting)과 과다적합 (overfitting)

underfitting과 overfitting 사이의 bias-variance tradeoff가 존재한다.



- **OLS**의 **underfitting** : 실제 beta는 k-dimensional vector여야 하는데, k-1 space에 projection 된다면 그 만큼의 편차가 발생할 수 밖에 없다. -> OLS는 더 이상 unbiased estimator가 아님.

- **OLS**의 **overfitting** : 주어진 종속변수를 설명하는데 X1으로 충분하지만 X2가 추가되었다고 하자. 이때 unbiaseness에는 문제가 없다. 그러나 $Var(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\sigma^2 I}(\boldsymbol{X^T X})^{-1}$이므로 variance가 증가하게 된다. 현재의 데이터셋을 잘 설명하지만 새로운 데이터가 들어왔을 때 예측력이 떨어진다.

- Overfitting을 해결하기 위한 방법 : (1) 쓸모 없는 **변수를 제거 (e.g. stepwise regression)**하거나 **차원을 축소 (e.g. PCA)**한다. (2) **데이터를 추가**한다 (bootstrapping, longer lookback window – <what if autocorrelation & regime change?>, simulation…) (3) 변수 추가에 대한 penalty를 부여한다 (**Regularization**)

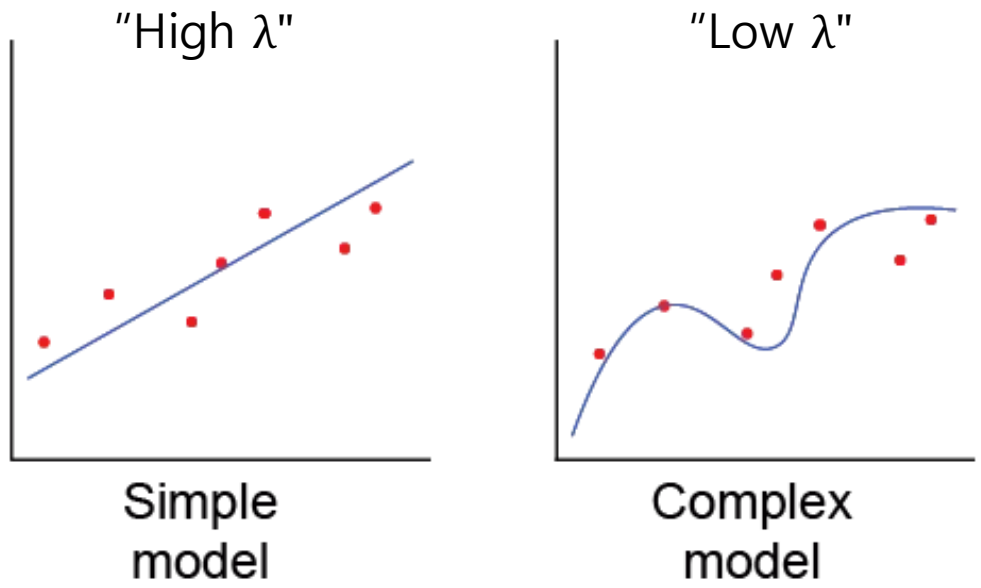
Regularization

Overfitting을 방지하기 위해 추가되는 독립변수에 대한 패널티를 준다.

$$L(x,y) = \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2$$

where $h_\theta x_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3 + \theta_4 x_4^4$

$$L(x,y) \equiv \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_i^2$$



- Least square 목적 함수를 변형. 변형된 목적 함수를 최소화 시키기 위해서는 estimator 값이 줄어들 필요가 있다.
- 이 때, 변형된 목적 함수의 미분값 $\frac{d}{d\boldsymbol{\beta}}\left[(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda \boldsymbol{I}\|\boldsymbol{\beta}\|^2\right] = \boldsymbol{0}$이 되도록 하는 $\boldsymbol{\beta}$를 구하면 된다.
- $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{y}$

고전적 가정에 위배될 때 일어나는 일

OLS는 더 이상 최고의 추정이 될 수 없으며 다른 모델을 사용해야 한다.

1. 이분산 (Heteroskedasticity) : OLS의 efficiency 깨지나 Unbiasnes는 유지됨. 원래의 독립변수 matrix를 약간 변형한 GLS를 사용하면 BLUE를 만들 수 있다. 또한 FGLS나 Robust inference 등의 방법이 있다.

2. 자기상관 (Autocorrelation) : 자기상관이 있으면 실제보다 오차를 작게 추정해 버림 (effective n < measured n) 따라서 유의하지 않은 변수를 유의하다고 잘못 결론내리게 됨. 잔차끼리 OLS를 해서 자기상관계수를 추정한 후 GLS 또는 FGLS를 하면 BLUE가 된다. 잔차의 자기상관을 처리하는 모형으로 AR, MA, ARMA, ARIMA등이 있다. 또한 시계열 데이터 잔차의 변동성(분산)자체에 자기상관이 존재하는 경우 ARCH, GARCH등의 모형을 쓰면 efficient해 진다.

3. 내생성 : $E(\boldsymbol{\varepsilon}|\boldsymbol{X})$가 더 이상 0이 아닌 경우이다. OLS 추정량의 불편성이 깨지고 consistency 또한 깨질 수 있다 (covariance가 0이 안되는 경우). 도구변수(instrument variable)를 도입하여 2SLS를 하거나 연립방정식 모형을 구축하여 해결한다.

4. Binary response : 종속변수와 독립변수간의 관계가 binary (0 or 1)와 같이 discrete한 경우 classification 문제가 된다. 이때는 binary response를 함수값으로 가지는 logit (logistic function의 역함수) 모형으로 logistic regression을 하는 경우가 일반적이다. 정규분포 누적분포함수의 역함수인 probit 모형이나 hypertangent 함수 또한 사용한다.

그 외 신경 써야 하는 사항들

.

다중공선성 : $\mathrm{VIF}_i = \frac{\sigma^2}{(n-1)\mathrm{Var}[X_i]} \cdot \frac{1}{1-R_i^2}$

레버리지 h : $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{iN}y_N$ 특정 데이터가 추정값에 영향을 주는 정도.

아웃라이어 : $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$ 스튜던트와 잔차가 크면 아웃라이어.

Cook's distance : $D_i = \frac{r_i^2}{\mathrm{RSS}}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$ $D_i > \frac{4}{N-K-1}$ 일 때 아웃라이어로 판단.

회귀분석을 이용한 퀀트 모델의 예시

- 수익률을 설명하는 독립적인 팩터를 이용하여 회귀분석 주식예측모델을 형성
- 높은 기대수익률을 가지는 종목에 가중을 둬서 매수, 낮은 기대수익률을 가지는 종목에 가중을 둬서 매도하는 전략 (momentum)
- 기대수익률에 비해 현재의 수익률이 표준편차 보다 작다면 매수, 표준편차보다 크다면 매도하는 전략 (mean –reversion)
- 거래비용 및 투자가의 제약조건 등을 감안하여 최적 포트폴리오를 구축
- 한 달 예측에 많이 사용되는 팩터 예시:

Factors	
Price-to-book ratio	Earnings Certainty
Gross profit / Total assets	Cash flow yield
ROE	Dividend yield
Net Margin	Realized vol
Asset Turnover	1M momentum
Gearing	12M-1M momentum
Forward earnings yield	Market cap