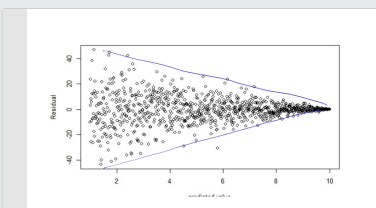
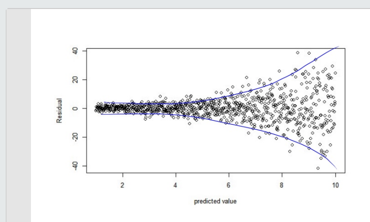
Workspace for 'SM_Lecture2_2'

Page 1 (row 1, column 1)

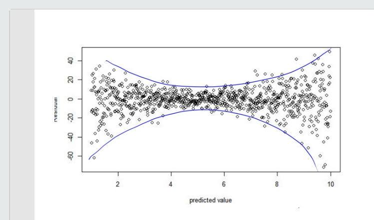


$\sigma^2 g(\cdot) = \sigma^2(\gamma_0 + \gamma_1 \hat{y}_i)$



$\sigma^2 g(\cdot) = \sigma^2[e^{\gamma \hat{y}_i}]$

$\sigma^2 g(\cdot) = \sigma^2[\gamma_0 + e^{\gamma_1 \hat{y}_i}]$



$\sigma^2 g(\cdot) = \sigma^2[\gamma_0 + \gamma_1 \hat{y}_i + \gamma_2 \hat{y}_i^2]$

$\hat{y}_i = f(X;\theta)$ 이므로 결국 $g(\cdot)$도 θ에 dependent 하다.

2. Statistical Modelling (2)

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

parametric한 모델링에 중점

$$y = \underset{\uparrow \atop \text{Specified}}{f}(x) + \varepsilon$$

STA3036

Data-Modelling Culture : parametric (small sample size, with information about the relationship between x and y)

Algorithmic Modelling Culture : non-parametric

Statistical Models

요즘 statistical Modelling이라고 한다면 Data-modelling과 Algorithmic-modelling을 합친 경우가 많으나, 수업시간엔 parametric한 경우만 알아보겠다.

- Statistical (data) model: A method to look at data / Summary (reduction) of data. Ex) $\underbrace{Y = \beta_0 + \beta_1 X}_{n\text{개의 데이터를 하나의 수식으로 축소한 것}}$

- Statistical models consist of two elements; systematic and random effects. Ex) ε Ex) β_i's
  - Systematic effects: Pattern of data. $\beta_0, \beta_1, \beta_2, \ldots$
  - Random effects: Unexplained or random variation. ε
  - Systematic effects are likely to be blurred by random effects.
  - Random effects are usually described in statistical terms.
    확률적 분포

- Looking intelligently at data $\Rightarrow$ Formulation of patterns $\Rightarrow$ Statistical data models.
  parametric
  - Succinct description of the systematic variation in the data.
  - Description patterns in similar data that might be collected for another study.

=> 분산/분포에 대한 간결한 설명

=> 유사한 데이터에 대한 설명력

Statistical Data Modelling (Parametric Models)

- E.g., consider the following model: $y = f(x; \theta)$.
  - No error & specified form of f.
  - For given $x_1, \ldots, x_n$, y takes the values $f(x_1; \theta), \ldots, f(x_n; \theta)$.
  - If θ is given, the values of y can be exactly reconstructed.
  - $\Rightarrow$ For given $x_1, \ldots, x_n$, θ is an exact summary of $y_1, \ldots, y_n$.

그러나 실제로

- Since there are errors in practice, the relationship between y and x has approximately f.
  - $\hat{y}_i = f(x_i; \hat{\theta})$, $i = 1, \ldots, n$: Theoretical or fitted values generated by the model f and the data.
  - The model cannot reproduce the original data values $y_1, \ldots, y_n$ exactly. ε가 확률분포를 따르기 때문에
  - The pattern from the model approximates the data values and it can be summarized by θ.

θ를 어떻게 추정할 것인가

- Estimation methods for data models:
  - => $n \to \infty$일 때, MLE는 consistent하고, minimum variance를 가지며, normality를 가진다.
  - Maximum likelihood estimation: Find model parameters maximizing the likelihood function for given data.
  - => θ가 fixed가 아닌 random variable로서 사용된다. (model-update에 용이하다) (불확실성을 수량화할 수 있다)
  - Bayesian estimation: Find the posterior distribution of model parameters for given prior distributions and likelihood function.
  - => 관찰되지 못한 불확실성에 대해 고려할 수 있다.
- This class focuses on the ML estimation.

Least Square Method

주된 machine learning 들의 가정 : $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Model: $Y = f(X; \boldsymbol{\theta}) + \epsilon$.
  - Y: Continuous variable.
  - f: Model.
  - $X = (X_1, \ldots, X_p)^\top$: Input variable vector.
  - $\boldsymbol{\theta}$ Model parameter vector.
  - ϵ: Random error.

- Least square method: Find $\boldsymbol{\theta}$ minimizing the discrepancy between y and $\hat{y}$.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

  where $\hat{y}_i = f(x_i; \hat{\boldsymbol{\theta}})$.
  - If (1) y_i's are statistically independent and (2) the variance of y_i does not depend on its mean value, the LS criterion is valid as a measure of discrepancy between y and $\hat{y}$.
  - Conditions (1) and (2) guarantee that all observations have the same weight. => universal variance를 갖고, independent 하기 때문에

Maximum Likelihood Estimation

- Data: $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$.
- Assumption: $X_1, \ldots, X_p$ are given (constant).
- Regression function: $E(Y|\underbrace{X = x}_{\text{어차피 constant 이니까}}) = f(x; \boldsymbol{\theta})$.
- Random variables in the data: $Y_1, \ldots, Y_n$.
- To construct the likelihood function, the joint distribution of $Y_1, \ldots, Y_n$, $p(\boldsymbol{Y}; \boldsymbol{\theta})$, should be identified.
- Likelihood function:

joint distribution of $Y_1, \ldots, Y_n$

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \equiv p(\boldsymbol{Y}; \boldsymbol{\theta}).$$

- MLE of model parameter $\boldsymbol{\theta}$: Let $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.
  - $\boldsymbol{\theta}$ maximizing $l(\boldsymbol{\theta})$ or $\boldsymbol{\theta}$ minimizing $\underbrace{-2l(\boldsymbol{\theta})}_{\text{deviance}}$.

Relationship between LS and ML

- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \; i = 1, \ldots, n.$
- $Y_i \overset{iid}{\sim} N(\mu_i, \sigma^2), \; i = 1, \ldots, n$, where $\mu_i = E(Y_i) = f(\boldsymbol{x}_i; \boldsymbol{\theta})$.
- Since Y_i's are independent, the joint density of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is

$\mu_i = f(x_i; \theta)$ 가 주어졌으므로

$$
\begin{aligned}
p(\boldsymbol{Y}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \mu_i)^2}{2\sigma^2} \right) \right] \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta}))^2}{2\sigma^2} \right) \right].
\end{aligned}
$$

Relationship between LS and ML

- MLE: For fixed σ^2,

$$
\begin{aligned}
\max_{\boldsymbol{\theta}}[l(\boldsymbol{\theta})] &\equiv \min_{\boldsymbol{\theta}}\left[-2l(\boldsymbol{\theta};\boldsymbol{y})\right] \\
&\equiv \min_{\boldsymbol{\theta}} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta}))^2 \\
&\equiv \min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta}))^2 \\
&\equiv \min_{\boldsymbol{\theta}}(\boldsymbol{y}-\boldsymbol{f})^\top(\boldsymbol{y}-\boldsymbol{f}) \\
&= \mathit{LS\ criterion},
\end{aligned}
$$

where $\boldsymbol{f} = (f(\boldsymbol{x}_1;\boldsymbol{\theta}),\ldots,f(\boldsymbol{x}_n;\boldsymbol{\theta}))^\top$.

$\therefore$ $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ 이 충족되면 MLE = LSE

When Error Assumptions are Violated

✱ 일반적인 선형회귀의 가정이 위배되었을 때 어떻게 모델링을 해야 하느냐
- Objective function을 Error Assumption이 위배되었을 때의 model로 modify해서 새로운 model을 유도

- Assumptions for error ϵ:
  (1) ϵ_i's have constant variance. Homoscedasticity
  (2) ϵ_i's are independent.
  (3) ϵ_i's have normal distribution.

  residual plot을 통해 간단히 확인가능

- From the residuals $r_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, we can check the assumptions (1), (2) and (3).

- When the assumptions are violated, the model variance $\uparrow \Rightarrow$ Poor prediction. ✱ n이 작을수록 assumptions violation은 치명적이다.

- How to solve these violations?
  - (1) $\Rightarrow$ Weighted least squares. (분산이 다르기 때문에 다른 가중치 부여)
  - (2) $\Rightarrow$ Covariance matrix (e.g., time/spatial). (Y_i's 들이 independent 하지 않을 때)
  - (3) $\Rightarrow$ Transformation. (Normality를 따르지 않을 때)

2주차 - 1 끝

Nonconstant Error

- Suppose that $\epsilon_i \sim N(0, \underbrace{\sigma_i^2}_{\text{heteroscedasticity}})$, $i = 1, \ldots, n$ and ϵ_i's are independent.

  ⇒ 등분산을 만족하지 못하면 $\hat{\beta}$의 분산이 커지게 되므로 모델의 신뢰도가 낮아짐

- Then, $\boldsymbol{Y} \sim MVN(\boldsymbol{f}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = diag\underbrace{(\sigma_1^2, \ldots, \sigma_n^2)}_{\text{non-constant variance}}$.
- Likelihood function:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{f}) \right\}.$$

- MLE: For known σ_i^2, $i = 1, \ldots, n$,

$$\begin{aligned} \max_{\boldsymbol{\theta}} [l(\boldsymbol{\theta})] &\equiv \min_{\boldsymbol{\theta}} \left[-2l(\boldsymbol{\theta}; \boldsymbol{y}) \right] \\ &\equiv \min_{\boldsymbol{\theta}} (\boldsymbol{y} - \boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{f}). \end{aligned}$$

Nonconstant Error

- Consider the linear regression model. That is, $\boldsymbol{f} = \boldsymbol{X}\beta$. Then MLE of β is given by

$\beta = (X'X)^{-1}X'Y$

pf) $\min_{\beta} Q(\beta) = \min_{\beta} (y - X\beta)^T \Sigma^{-1} (y - X\beta)$

$= \min_{\beta} y^T \Sigma^{-1} y - 2\beta^T X^T \Sigma^{-1} y + \beta^T X^T \Sigma^{-1} X \beta$

$\Rightarrow \frac{\partial}{\partial \beta} Q(\beta) \Big|_{\beta = \hat{\beta}} = -2X^T \Sigma^{-1} y + 2X^T \Sigma^{-1} X \hat{\beta} = 0$

$\hat{\beta} = \underbrace{(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y}_{\text{Weighted Least Square Estimation}}$

* Σ^{-1} can be called a "precision matrix"

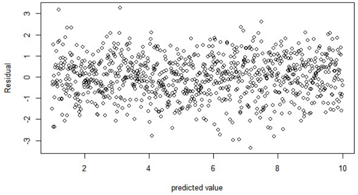
* Σ^{-1}의 각 element들이 weight가 된다 (σ_i^2 가 클수록 weight가 작아지고, σ_i^2 가 작을수록 weight가 커진다.)

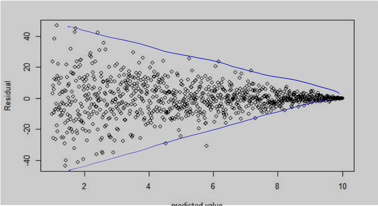
- MLE of β is the weighted least square estimator (WLSE).

Nonconstant Error with Pattern

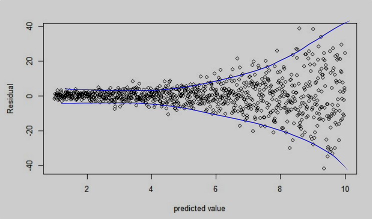
- If a residual plot shows some pattern, the variance function can be considered.

Irreducible Errors

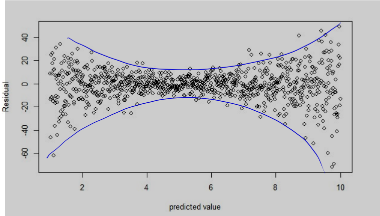




=> Suggested transformation : coefficients with negative sign



Suggested Transformation : exponential transformation



Suggested transformation : Quadratic Transformation

Variance Function

- Variance function: $Var(\epsilon_i) = \sigma^2 g^2(\boldsymbol{z}_i; \boldsymbol{\theta}, \gamma)$.
  - z_i: Known vector, possibly $\boldsymbol{x}_i$. z_i can be $\hat{y}_i, x_j, \underset{\sim}{x}$
  - σ: Unknown scale parameter.
  - $g(\cdot)$: Function to be estimated by parametric or nonparametric method.
  - $\boldsymbol{\theta}$: Parameter vector of the model f.
  - $\boldsymbol{\gamma}$: Parameter vector of the variance function.

Variance Function이 mean에 의존하는 경우도 있는데, 이러한 경우는 결국 variance가 model f에 의존한다는 것이다.
θ가 만약 mean에 의존하지 않는다면 variance function에서 빠져나올 수 있다.

- $Y_i \sim^{indep.} N(f(\boldsymbol{x}_i; \boldsymbol{\theta}),\ \sigma^2 g^2(\boldsymbol{z}_i; \boldsymbol{\theta}, \boldsymbol{\gamma})),\ i = 1, \ldots, n$.

  $y_i = f(\underset{\sim}{x}_i ; \theta) + \sigma g(\cdot) \cdot \eta_i$, η_i : pure error term $\sim N(0,1)$
- Examples of variance function:
  - Linear pattern: $\sigma g(\boldsymbol{z}_i; \boldsymbol{\theta}, \gamma) = \boldsymbol{z}_i^\top \gamma$.
  - Exponential pattern: $\sigma^2 g^2(\boldsymbol{z}_i; \boldsymbol{\theta}, \gamma) = \exp(\boldsymbol{z}_i^\top \gamma)$.

- $Var(Y_i)$ often depends on its mean $E(Y_i)$. In that case, $\boldsymbol{z}_i$ can be replaced with $\hat{y}_i = f(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}})$.

Variance Function Estimation

Variance parameters are usually nuisance parameters

- Log likelihood function:

$$\max_{\boldsymbol{\theta},\gamma,\sigma} l(\boldsymbol{\theta},\gamma,\sigma;\boldsymbol{y},\boldsymbol{z}) = \max_{\boldsymbol{\theta},\gamma,\sigma} -\sum_{i=1}^{n}\log\{\sigma g(\boldsymbol{z}_i;\boldsymbol{\theta},\gamma)\} - \frac{1}{2}\sum_{i=1}^{n}\left\{\frac{(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta}))^2}{\sigma^2 g^2(\boldsymbol{z}_i;\boldsymbol{\theta},\gamma)}\right\}.$$

- In this maximization problem, it is not easy to find $\boldsymbol{\theta},\gamma,\sigma$ simultaneously.

- Pseudolikelihood estimation:
  - To find γ and σ, it maximizes $l(\hat{\boldsymbol{\theta}},\gamma,\sigma;\boldsymbol{y},\boldsymbol{z})$, where $\hat{\boldsymbol{\theta}}$ is the MLE from $l(\boldsymbol{\theta},\hat{\gamma},\hat{\sigma};\boldsymbol{y},\boldsymbol{z})$.
  - Estimations of $\boldsymbol{\theta}$ and (γ,σ) are iterated until $\hat{\boldsymbol{\theta}}$ is converged.

Variance Function Estimation

증명은 복잡해서 pass !

- Residual: $r_i = y_i - f(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}})$.

- $E(r_i^2) \approx \sigma^2 g^2(\boldsymbol{z}_i; \boldsymbol{\theta}, \gamma)$.

- If ϵ_i's have normal distribution, $Var(r_i^2) \approx \sigma^4 g^4(\boldsymbol{z}_i; \boldsymbol{\theta}, \gamma)$.

  proof is beyond the scope of this course

- Weighted estimator: γ and σ minimizing

$$\sum_{i=1}^{n} \frac{[r_i^2 - \sigma^2 g(\boldsymbol{z}_i; \gamma, \boldsymbol{\theta})]^2}{\sigma^4 g^4(\boldsymbol{z}_i; \gamma, \boldsymbol{\theta})}.$$

Generalized least squares

Algorithm

When residuals have patterns

1. Set the initial parameter vectors $\hat{\boldsymbol{\theta}}$, $\hat{\gamma}$, $\hat{\sigma}$.
2. For given $\hat{\boldsymbol{\theta}}$, compute squared residuals $r_i^2 = [y_i - f(x_i; \hat{\boldsymbol{\theta}})]^2$.
3. Estimate the variance function parameters γ and σ by minimizing
   by Profile-method

$$\min_{\gamma,\sigma} \sum_{i=1}^{n} \frac{[r_i^2 - \sigma^2 g(\boldsymbol{z}_i; \gamma, \hat{\boldsymbol{\theta}})]^2}{\hat{\sigma}^4 g^4(\boldsymbol{z}_i; \hat{\gamma}, \hat{\boldsymbol{\theta}})}.$$

4. Estimate $\boldsymbol{\theta}$ maximizing $l(\boldsymbol{\theta}, \hat{\gamma}, \hat{\sigma}; \boldsymbol{y}, \boldsymbol{z})$.
5. Iterate Steps 2–4 until $\boldsymbol{\theta}$ is converged.