2. Statistical Modelling (2)

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Statistical Models

- Statistical (data) model: A method to look at data / Summary (reduction) of data.

- Statistical models consist of two elements; systematic and random effects.
  - Systematic effects: Pattern of data.
  - Random effects: Unexplained or random variation.
  - Systematic effects are likely to be blurred by random effects.
  - Random effects are usually described in statistical terms.

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

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 $y_1, \ldots, y_n$.

- 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.
  - The pattern from the model approximates the data values and it can be summarized by θ.

Fitting Data Models

- Estimation methods for data models:
  - Maximum likelihood estimation: Find model parameters maximizing the likelihood function for given data.
  - Bayesian estimation: Find the posterior distribution of model parameters for given prior distributions and likelihood function.
- This class focuses on the ML estimation.

Least Square Method

- 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.

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: $E(Y|X = x) = f(x; \boldsymbol{\theta})$.
- 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:

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

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

Relationship between LS and ML

- $\epsilon_i \ \sim^{iid} \ N(0, \sigma^2), \ i = 1, \ldots, n.$
- $Y_i \ \sim \ N(\mu_i, \sigma^2), \ i = 1, \ldots, n$, 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

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

Relationship between LS and ML

- MLE: For fixed σ^2,

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

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

When Error Assumptions are Violated

- Assumptions for error $\boldsymbol{\epsilon}$:
  1. ϵ_i's have constant variance.
  2. ϵ_i's are independent.
  3. ϵ_i's have normal distribution.

- 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.

- How to solve these violations?
  - (1) $\Rightarrow$ Weighted least squares.
  - (2) $\Rightarrow$ Covariance matrix (e.g., time/spatial).
  - (3) $\Rightarrow$ Transformation.

Nonconstant Error

- Suppose that $\epsilon_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, n$ and ϵ_i's are independent.
- Then, $\boldsymbol{Y} \sim MVN(\boldsymbol{f}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$.
- 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, $i = 1, \ldots, n$,

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

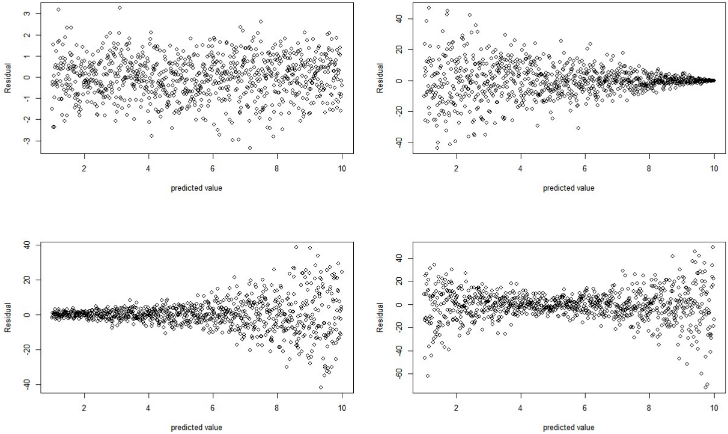
Nonconstant Error

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

- 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.



Variance Function

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

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

- 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

- 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

- 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, $Var(r_i^2) \approx \sigma^4 g^4(\boldsymbol{z}_i; \boldsymbol{\theta}, \gamma)$.
- Weighted estimator: γ and σ minimizing

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

Algorithm

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

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.