Regression Analysis Tutoring3

Seung Bong Jung

Seoul National University

November 1, 2021

Variance stabilizing transformations

- In the usual linear regression, we assume the homoskedasticity of the error term. However, the given data may violate this assumption because the variance is functionally related to the mean.(heteroskedasticity).
- We may stabilize the variance by taking transformation on the response(to achieve homoskedasticity).

Relationship of σ^2 to $\mathsf{E}(y)$	Transformation
$\sigma^2 \propto \mathsf{E}(y)$	$y' = \sqrt{y}$
$\sigma^2 \propto \mathsf{E}(y)(1 - \mathsf{E}(y))$	$y' = \arcsin\sqrt{y}$
$\sigma^2 \propto \mathsf{E}(y)^2$	$y' = \log y$
$\sigma^2 \propto \mathsf{E}(y)^3$	$y' = y^{-\frac{1}{2}}$
$\sigma^2 \propto \mathsf{E}(y)^4$	$y' = y^{-1}$

Table: Variance stabilizing transformation

Box-Cox method

- If the distribution of the given data does not satisfy normality, one can take transformation on the data using Box-Cox method to attain normality.
- But one should note that Box-Cox method does not guarantee the normality. It only assumes that among all transformations with Lambda values between -5 and $+5$, transformed data has the highest likelihood, making variance smaller. However, the smallest variance does not ensure the normality. One has to check whether the transformed data follows normal distribution.(e.g. Kolmogorov-Smirnov test, Q-Q plot, P-P plot)
- Note the response should be positive.

Box-Cox method

- $y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda \check{y}^{\lambda-1}}$ if $\lambda \neq 0$, and $y^{(\lambda)} = \check{y} \log y$, otherwise, where $\check{y} = \log^{-1}(\frac{1}{n}\sum_{i=1}^n \log y_i)$
- With $y^{(\lambda)}$, we consider the model $\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Note we're estimating $\theta = (\boldsymbol{\beta}, \lambda)$. Under the normality for $\boldsymbol{\epsilon}$, we estimate θ by maximizing the likelihood, equivalently minimizing

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} (y_i^{(\lambda)} - \mathbf{X}_i \boldsymbol{\beta})^2.$$

Transformation on response variable

- One can linearize the nonlinear model, by taking appropriate transformation.
- For example, if $y = \beta_0 x^{\beta_1}$, one can take log transformation. Also, if $y = \frac{x}{\beta_0 x - \beta_1}$. one can take reciprocal transformation.
- Generally, consider the following nonlinear model:

$$y = f(\alpha, \beta_0, \beta_1) + \epsilon = \beta_0 + \beta_1 \xi + \epsilon$$

,where $\xi = x^\alpha$ if $\alpha \neq 0$, and $\xi = \log x$ if $\alpha = 0$.

- Take initial value $\alpha_0 = 1$. Viewing $f(\alpha, \beta_0, \beta_1)$ as a function of α with β_0, β_1 being fixed, by Taylor's expansion, we have

$$\begin{aligned} f(\alpha, \beta_0, \beta_1) &\approx f(\alpha_0, \beta_0, \beta_1) + (\alpha - \alpha_0)\{\frac{\partial f(\alpha, \beta_0, \beta_1)}{\partial \alpha}\}|_{\alpha=\alpha_0} \\ &= \beta_0 + \beta_1 x + (\alpha - 1)\beta_1 (x^\alpha \log x)|_{\alpha=1} \\ &= \beta_0 + \beta_1 x + (\alpha - 1)\beta_1 x \log x \end{aligned}$$

Transformation on response variable

- Hence, $f(\alpha, \beta_0, \beta_1) \approx \beta_0 + \beta_1 x + \gamma w$, where $w = x \log x$. Then clearly, $\hat{\alpha} = 1 + \frac{\hat{\gamma}}{\hat{\beta}_1}$.
- Since the MLE of $\theta = (\alpha, \beta_0, \beta_1)$ is not always can be obtained in closed form, one has to resort to iterative method.
- Under the assumption $y = f(x, \theta) + \epsilon$, where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, MLE $\hat{\theta}$ of θ is $\arg\min \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2$. Let $S(\theta) = \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2$.
- Finding $\hat{\theta}$ can be equivalent to solving the equation $\frac{\partial S(\theta)}{\partial \theta} = 0$, under some appropriate conditions.

Transformation on the response variable

- $\frac{\partial S(\theta)}{\partial \theta} \propto U(\theta) = \sum_{i=1}^{n} z_i(y_i - f(x_i, \theta))$, where $z_i = \frac{\partial f(x_i,\theta)}{\partial \theta}$
- The solution can be found using Newton method. With initial value θ_0, update $\theta^{(p)}$ using the following until the stopping criterion is satisfied :

$$\theta^{(p+1)} = \theta^{(p)} + I(\theta^{(p)})^{-1} U(\theta^{(p)})$$

  where $I(\theta) = -\frac{\partial^2 U(\theta)}{\partial \theta \partial \theta^t}$.
- Alternatively, one can use Fisher's Scoring method, which uses $\mathsf{E}(I(\theta))$ instead of $I(\theta)$,
- Note as $n \to \infty$, $\hat{\theta} \xrightarrow{d} \mathcal{N}_p(\theta_0, \sigma^2 Z(\theta_0) Z(\theta_0)^t)$, where $Z(\theta)$ is $p \times p$ matrix with ith row being z_i.
- Instead of Newton method, one can try gradient descent method with step size γ, whose update rule is

$$\theta^{(p+1)} = \theta^{(p)} - \gamma \nabla S(\theta^{(p)})$$

Generalized Least Squared Estimation

- For the random vector Y, assume that $\mathrm{E}(Y) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Var}(Y) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is positive-definite matrix and known.
- Then the least squared estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is

$$(\mathbf{X}^{\mathbf{t}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{t}\mathbf{V}^{-1}\mathbf{Y}$$

- Note that the least squared estimator $\hat{\boldsymbol{\beta}}$ is BLUE.