Chapter 5. Transformation and Weighting to Correct Model Inadequacy

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

One may consider making transformation on the response variable if some of the basic assumptions are invalid. Practically, decide on a transformation, go ahead to fit it then examine the residuals in the metric of the *transformed* variable.

e.g. Check $f(y_i) - \widehat{f(y_i)}$ in $f(y) = X\beta + \epsilon$ for normality, etc.

Also all tests and confidence statements must be made in the *transformed* space. e.g. $f(y)$ is predicted by $\widehat{f(y)}$, then $\hat{y} = f^{-1}(\widehat{f(y)})$.

Note: 1. Usually, no mathematical equivalence between parameters of the two models.

2. There is **no** formal analysis to find a transformation. Inoformal plots of the data will reveal the need for some 'special' and 'well known' transformations.

Variance-Stabilizing Transformation

Q: When is $Var(\epsilon_i) \neq \sigma^2$ (from e_i vs. $\hat{y}_i$ plots)?

Ans: $Var(y)$ may depend on $E(y)$, e.g. in Poisson distribution,

$$Var(y) = E(y).$$

Let $\eta = E(y)$, $\sigma_y = (Var(y))^{1/2} = g(\eta)$. We want to find a transformation $y' = h(y)$ such that $Var(y') =$ constant.

Idea: (Delta method.) $y \sim N(\eta, \sigma_y^2)$, then

$$y' = h(y) \sim N(h(\eta), \sigma_y^2 (h'(\eta))^2).$$

Thus

$$Var(y') = \sigma_y^2 (h'(\eta))^2 = (g(\eta)h'(\eta))^2 = \text{ constant}.$$

$\therefore$ Try to find h such that

$$h'(y) \propto \frac{1}{g(y)} \quad \text{or} \quad h(y) = \int (g(y))^{-1} dy.$$

e.g.: 1) $\sigma_y \propto$ constant. $y' = y$.

2) $\sigma_y \propto \sqrt{E(y)}$ $(\sigma_y \propto \sqrt{\eta}), y \geq 0$ (Poisson, e.g.),

$y' = h(y) = \sqrt{y}$.

3) $\sigma_y \propto E(y) = \eta, y > 0$, $y' = \ln y$.

4) $\sigma_y \propto (E(y))^2 = \eta^2$, $y' = 1/y$.

5) $\sigma_y \propto \sqrt{E(y)(1 - E(y))} (= \sqrt{\eta(1-\eta)})$,

$(0 \leq y \leq 1, ny \sim$ binomial), $y' = \sin^{-1} \sqrt{y}$.

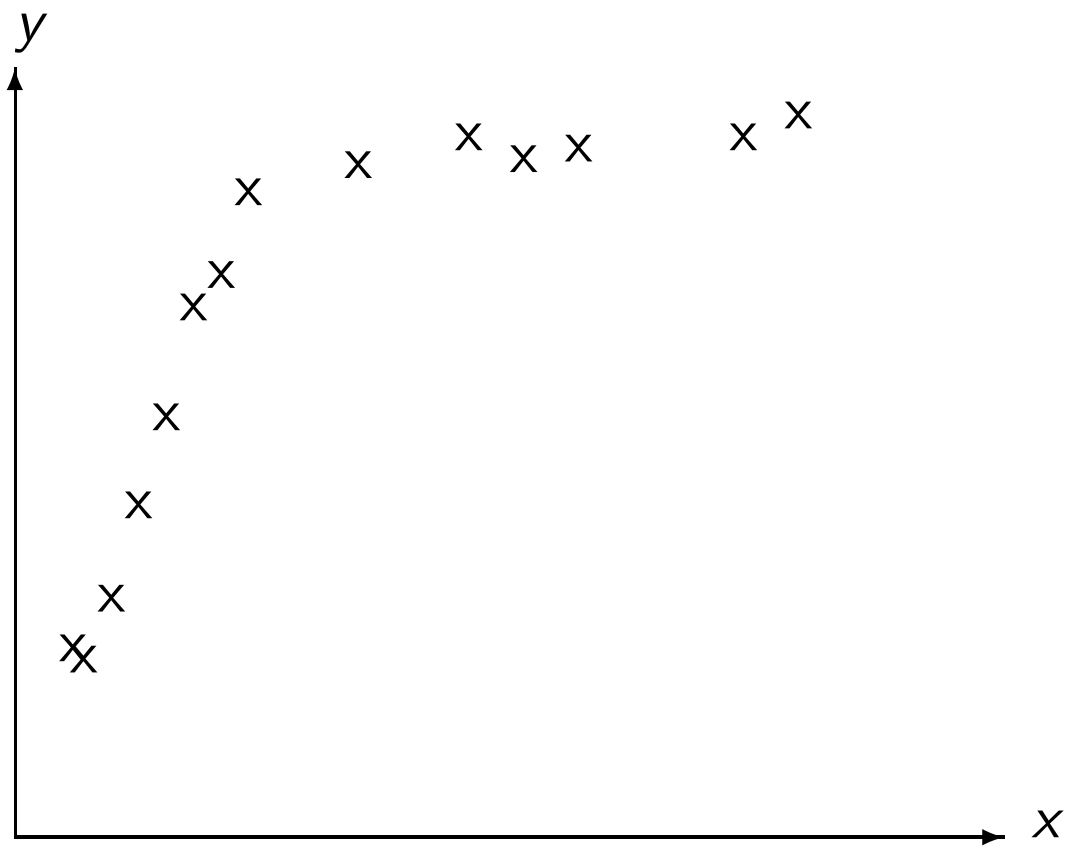
$((\sin^{-1} y)' = 1/\sqrt{1 - y^2})$. □

Linearize the Model

In y_i vs. x_i plots, certain curvatures are presented. One may consider to use some '*linearizable*' functions.

See Table 5.4 on p. 165.

e.g.



May consider (1) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ or

(2) ** $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + \epsilon.$

Power Transformation: Box-Cox Method

For $y > 0$, consider $y' = y^{\lambda}$. $\lambda =?$ MLE.

If $y_{(n)}/y_{(1)}$ is considerably large (*long-tail*), we may consider a transformation

$$z = y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

Then $y^{(\lambda)}$ is continuous and differentiable in λ.

Consider

$$\mathbf{z} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I).$$

Then the likelihood function of $(\lambda, \boldsymbol{\beta}, \sigma^2)$ is

$$l(\lambda, \boldsymbol{\beta}, \sigma^2 | \mathbf{z}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\{-\frac{1}{2\sigma^2}(\mathbf{z} - X\boldsymbol{\beta})'(\mathbf{z} - X\boldsymbol{\beta})\},$$

based on the *transformed* data $z_1, \ldots, z_n$.

Note that the likelihood function based on the original data $y_1, \ldots, y_n$ is

$$\begin{aligned} l_{\mathbf{y}}(\lambda, \boldsymbol{\beta}, \sigma^2) &= f(\mathbf{y}|\lambda, \boldsymbol{\beta}, \sigma^2) \\ &\propto (\sigma^2)^{-n/2} \\ &\quad \times \exp\{-\frac{1}{2\sigma^2}(\mathbf{y}^{(\lambda)} - X\boldsymbol{\beta})'(\mathbf{y}^{(\lambda)} - X\boldsymbol{\beta})\}|J(\lambda)|, \end{aligned}$$

where $J(\lambda) = \left(\frac{\partial z_i}{\partial y_j}\right) = \left(\frac{\partial y_i^{(\lambda)}}{\partial y_j}\right) = diag\left(\frac{\partial y_i^{(\lambda)}}{\partial y_i}\right)$.

Now

$$\frac{dy^{(\lambda)}}{dy} = \begin{cases} y^{\lambda-1}, & \lambda \neq 0 \\ 1/y, & \lambda = 0 \end{cases} . \quad \therefore |J(\lambda)| = \begin{cases} \prod_{i=1}^{n} y_i^{\lambda-1}, & \lambda \neq 0 \\ \prod_{i=1}^{n} y_i^{-1}, & \lambda = 0. \end{cases}$$

$$\begin{aligned} \mathcal{L}_{\mathbf{y}}(\lambda, \boldsymbol{\beta}, \sigma^2) &= \log l_{\mathbf{y}}(\lambda, \boldsymbol{\beta}, \sigma^2) \\ &\propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}^{(\lambda)} - X\boldsymbol{\beta})'(\mathbf{y}^{(\lambda)} - X\boldsymbol{\beta}) + \log|J(\lambda)| \end{aligned}$$

Then maximize $\mathcal{L}_{\mathbf{y}}(\lambda, \boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}, \sigma^2$ for given λ.

For fixed λ, $\mathcal{L}_{\mathbf{y}}(\lambda, \boldsymbol{\beta}, \sigma^2)$ is maximized at

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}^{(\lambda)} \quad \text{and} \quad \hat{\sigma^2} = (\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}})'(\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}})/n,$$

and

$$\begin{aligned}
\mathcal{L}_{\mathbf{y}}(\lambda, \hat{\boldsymbol{\beta}}, \hat{\sigma^2}) &= -\frac{n}{2}\log(\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}})'(\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}})/n \\
&\quad -(\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}})'(\mathbf{y}^{(\lambda)} - X\hat{\boldsymbol{\beta}})/(2\hat{\sigma^2}) + \log|J(\lambda)| \\
&= -\frac{n}{2}\log \mathbf{y}^{(\lambda)'}(I - H)\mathbf{y}^{(\lambda)}/n - \frac{n}{2} + \frac{n}{2}\log|J(\lambda)|^{2/n}.
\end{aligned}$$

Finally, we need to find λ such that $\mathcal{L}_{max}(\lambda) = \mathcal{L}_{\mathbf{y}}(\lambda, \hat{\boldsymbol{\beta}}, \hat{\sigma^2})$ is maximized. Note that

$$\begin{aligned}
\mathcal{L}_{max}(\lambda) &= \begin{cases} -\frac{n}{2}\log \frac{\mathbf{y}^{(\lambda)'}(I-H)\mathbf{y}^{(\lambda)}}{n\left(\prod_{i=1}^n y_i^{\lambda-1}\right)^{2/n}}, & \lambda \neq 0 \\ -\frac{n}{2}\log \frac{(\ln \mathbf{y})'(I-H)(\ln \mathbf{y})}{n\left(\prod_{i=1}^n y_i^{-1}\right)^{2/n}}, & \lambda = 0 \end{cases} \\
&= \begin{cases} -\frac{n}{2}\log \frac{\mathbf{y}^{(\lambda)'}(I-H)\mathbf{y}^{(\lambda)}}{n\left[\left(\prod_{i=1}^n y_i\right)^{1/n}\right]^{2(\lambda-1)}}, & \lambda \neq 0 \\ -\frac{n}{2}\log \frac{(\ln \mathbf{y})'(I-H)(\ln \mathbf{y})}{n\left[\left(\prod_{i=1}^n y_i\right)^{1/n}\right]^{-2}}, & \lambda = 0 \end{cases}
\end{aligned}$$

Let

$$v_i = \begin{cases} \mathbf{y}_i^{(\lambda)}/(\dot{\mathbf{y}})^{\lambda-1} = \dfrac{\mathbf{y}_i^{\lambda}-1}{\lambda \dot{\mathbf{y}}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{\mathbf{y}} \ln y_i, & \lambda = 0, \end{cases}$$

where $\dot{\mathbf{y}} = (\prod_{i=1}^{n} y_i)^{1/n}$, *geometric mean* of $y_1, \ldots, y_n$. Hence,

$$\mathcal{L}_{max}(\lambda) = -\frac{n}{2} \log \frac{\mathbf{v}'(I-H)\mathbf{v}}{n},$$

and it is maximized as $\mathbf{v}'(I-H)\mathbf{v}$ is minimized.

Thus, to find the MLE of λ is the same as to find the λ such that $SS_{Res}(\lambda) = \mathbf{v}'(I-H)\mathbf{v}$ is minimized based on $\mathbf{v} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

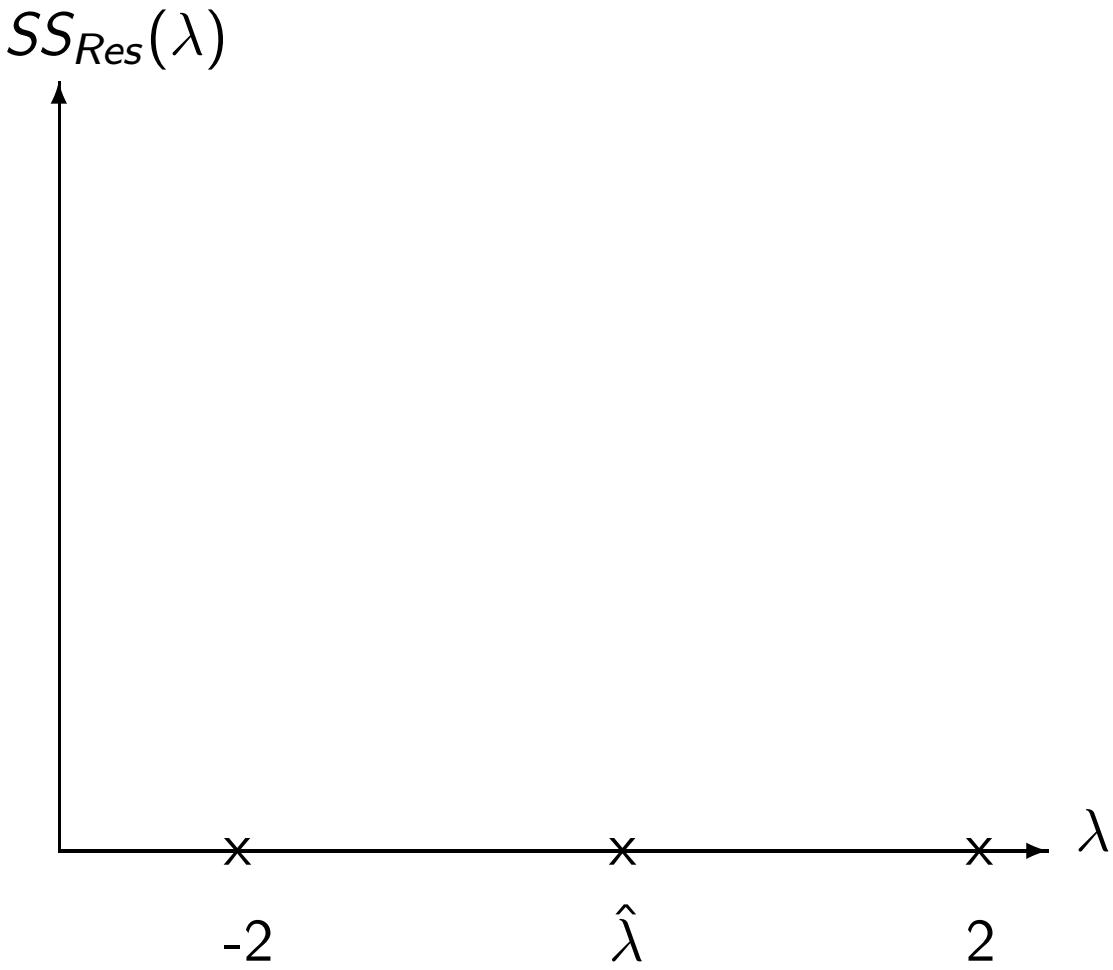
$\therefore$ The appropriate procedure is to use

$$\mathbf{v} = \begin{cases} \dfrac{\mathbf{y}^{\lambda}-1}{\lambda \dot{\mathbf{y}}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{\mathbf{y}} \ln \mathbf{y}, & \lambda = 0. \end{cases}$$

and

$$\min_{\lambda} SS_{Res}(\lambda) = \min_{\lambda} \mathbf{v}'(I - H)\mathbf{v}.$$

Usually, choose $\lambda \in (-2, 2)$ and plot $SS_{Res}(\lambda)$ versus λ.


$SS_{Res}(\lambda)$
λ
-2
$\hat{\lambda}$
2

Interval Estimates of λ

For large n, $\mathcal{L}_{max}(\hat{\lambda}) - \mathcal{L}_{max}(\lambda) \xrightarrow{d} \frac{1}{2}\chi^2_1$.

$\therefore$ an approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\begin{aligned} & \{\lambda : \ \mathcal{L}_{max}(\hat{\lambda}) - \mathcal{L}_{max}(\lambda) \leq \frac{1}{2}\chi^2_{1;\alpha}\} \\ = \quad & \{\lambda : \ \mathcal{L}_{max}(\lambda) \geq \mathcal{L}_{max}(\hat{\lambda}) - \frac{1}{2}\chi^2_{1;\alpha}\}. \end{aligned}$$

Or find λ such that

$$\begin{aligned} & \log \frac{SS_{Res}(\lambda)}{n} \leq \log \frac{SS_{Res}(\hat{\lambda})}{n} + \frac{\chi^2_{1;\alpha}}{n} \\ \iff \quad & SS_{Res}(\lambda) \leq SS_{Res}(\hat{\lambda}) e^{\chi^2_{1;\alpha}/n}. \end{aligned}$$

Note: 1. $e^x \approx 1 + x$ if x is small.

2. $\chi^2_1 = z^2 \approx t_\nu$, large $\nu (= n - p)$. $\therefore$ $e^{\chi^2_{1;\alpha}/n}$ can be replaced by

(i) $1 + t^2_{\nu:\alpha/2}/\nu$;

(ii) $1 + z^2(\alpha/2)/\nu$;

(iii) $1 + \chi^2_{1;\alpha}/\nu$;

(iv) $1 + \chi^2_{1;\alpha}/n$;

(v) $1 + z^2(\alpha/2)/n$.

Generalized and Weighted Least Squares

In $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, if $E(\boldsymbol{\epsilon}) = 0$, $var(\boldsymbol{\epsilon}) = \sigma^2 V$, where V is nonsigular, symmetric and posiitively definite. Then there exists a nonsigular, symmetric matrix $K_{n\times n}$ such that $V = K'K = KK$.

Def: The matrix K is often called the **square root** of V.

Note that $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is same as $K^{-1}\mathbf{y} = K^{-1}X\boldsymbol{\beta} + K^{-1}\boldsymbol{\epsilon}$.

Let $\mathbf{z} = K^{-1}\mathbf{y}$, $B = K^{-1}X$ and $\mathbf{g} = K^{-1}\boldsymbol{\epsilon}$, then it is equivalent to

$$\mathbf{z} = B\boldsymbol{\beta} + \mathbf{g},$$

where $E(\mathbf{g}) = K^{-1}E(\boldsymbol{\epsilon}) = 0$, and

$$Var(\mathbf{g}) = K^{-1}Var(\boldsymbol{\epsilon})(K^{-1})' = \sigma^2 K^{-1}VK^{-1} = \sigma^2 K^{-1}(KK)K^{-1} = \sigma^2 I.$$

So, the LSE of $\boldsymbol{\beta}$ is the one minimizing $S(\boldsymbol{\beta}) = \sum_{i=1}^{n} g_i^2 = \mathbf{g}'\mathbf{g}$ or

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (B'B)^{-1}B'\mathbf{z} = (X'K^{-1}K^{-1}X)^{-1}X'K^{-1}K^{-1}\mathbf{y} \\
&= (X'(KK)^{-1}X)^{-1}X'(KK)^{-1}\mathbf{y} \\
&= (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}, \quad \text{Generalized LSE of } \boldsymbol{\beta}.
\end{aligned}$$

Note:

1. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
2. $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(X'V^{-1}X)^{-1}$.
3. $SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'B'\mathbf{z} = \mathbf{y}'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$,

   $df = p$.
4. $SS_{Res} = \mathbf{z}'\mathbf{z} - \hat{\boldsymbol{\beta}}'B'\mathbf{z}$

   $= \mathbf{y}'V^{-1}\mathbf{y} - \mathbf{y}'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$,

   $df = n - p$.
5. $\mathbf{z}'\mathbf{z} = \mathbf{y}'V^{-1}\mathbf{y}$, $df = n$.

Special Case: $V = \begin{bmatrix} 1/w_1 & \cdots & \cdots & 0 \\ 0 & 1/w_2 & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 1/w_n \end{bmatrix}$, i.e. ϵ_i uncorrelated, but with unequal variances.

Let $W = V^{-1} = diag(w_i)$, then

$\hat{\boldsymbol{\beta}} = (X'WX)^{-1}X'W\mathbf{y} \equiv$ solution of min. weighted least squares,

which minimizes

$$\mathbf{g}'\mathbf{g} = \boldsymbol{\epsilon}' K^{-1} K^{-1} \boldsymbol{\epsilon} = \boldsymbol{\epsilon} V^{-1} \boldsymbol{\epsilon} = \sum_{i=1}^{n} w_i \epsilon_i^2.$$

i.e. $\min \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$. $\therefore \hat{\boldsymbol{\beta}}$ is also called the **weighted LSE**, WLSE.

Note: 1. w_i must be known or gained from prior knowledge.

2. Sometimes, $w_i = 1/x_i$, or $Var(\epsilon_i) \propto x_i$ revealed from the residual plots.

3. The **OLSE** (ordinary LSE $= (X'X)^{-1}X'\mathbf{y}$) is unbiased, but no longer a minimum variance estimator if $V \neq I$, for

$$\begin{aligned} Var(\hat{\beta}_{OLSE}) &= (X'X)^{-1}X'Var(\mathbf{y})X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'VX(X'X)^{-1}, \end{aligned}$$

and $Var(\hat{\beta}_{WLSE}) = \sigma^2(X'V^{-1}X)^{-1}$.

Note: In general, V is unknown and therefore must be estimated. There are $n(n+1)/2$ distinct elements in V, it would be impossible to reliably estimate all of them on the basis of n observations. However, if there exist known relationships involving very few parameters in V, then estimation precedures becomes available.

In the diagonal case, $V = diag(\sigma_i^2)$, an estimator of $\boldsymbol{\sigma}^{(2)} = (\sigma_1^2, \cdots, \sigma_n^2)$ is $\widehat{\boldsymbol{\sigma}^{(2)}} = (\hat{\sigma^2}_1, \cdots, \hat{\sigma^2}_n)'$ such that

$$\mathbf{e}^{(2)} = \begin{pmatrix} e_1^2 \\ \vdots \\ e_n^2 \end{pmatrix} = M^{(2)} \widehat{\boldsymbol{\sigma}^{(2)}},$$

where $M^{(2)} = (m_{ij}^2)$ with $I - H = (m_{ij})$.

Idea: $\because \ \mathbf{e} = (I - H)\boldsymbol{\epsilon}$, i.e. $e_i = \sum_{j=1}^{n} m_{ij}\epsilon_j,\ i = 1, \ldots, n,$

$$\therefore \ e_i^2 = \sum_{j=1}^{n}\sum_{l=1}^{n} m_{ij}m_{il}\epsilon_j\epsilon_l \quad \text{and} \quad E(e_i^2) = \sum_{j=1}^{n} m_{ij}^2 E(\epsilon_j^2),\ i = 1, \ldots, n.$$

Hence, σ_i^2 can be estimated by replacing $E(e_i^2)$ by e_i^2.

Note: These estimators are known as to be *MINQUE* (Minimized norm Quadratic Unbiased estimator). A major problem is that $\widehat{\sigma_i^2}$ can be negative.

Homework 6: (Page 185) 5.2, 5.7, 5.10, 5.12, 5.13, 5.14, 5.15, 5.16.

Due: Dec. 12, 2008.