Generalized Least Squares (GLS)

- What if the errors are not iid? E.g., $\mathbf{e} \sim \mathsf{N}_n(\mathbf{0}, \Sigma)$.
- Σ known (an ideal case for us to get some insight).
- Σ unknown (e.g., regression with time series data).
- Examples and R code.

GLS: Σ Known

- Assume $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ and $\mathbf{e} \sim \mathsf{N}_n(\mathbf{0}, \Sigma)$.

- Transform this problem back to OLS. Write $\Sigma = SS^T$ where we assume S^{-1} exists, then

$$
\begin{aligned}
S^{-1}\mathbf{y} &= S^{-1}\Big(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}\Big) \\
\mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^* \\
\mathbf{e}^* &\sim \mathsf{N}\Big(S^{-1}\mathbf{0}, S^{-1}\Sigma(S^{-1})^T\Big) = \mathsf{N}\Big(\mathbf{0}, \mathbf{I}\Big).
\end{aligned}
$$

- Now we can solve $\boldsymbol{\beta}$ using OLS,

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^*, \quad \mathbf{y}^* = S^{-1}\mathbf{y},\ \mathbf{X}^* = S^{-1}\mathbf{X}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left[(\mathbf{X}^*)^T\mathbf{X}^*\right]^{-1}(\mathbf{X}^*)^T\mathbf{y}^* \\ &= (\mathbf{X}^T(S^{-1})^TS^{-1}\mathbf{X})^{-1}\mathbf{X}^T(S^{-1})^TS^{-1}\mathbf{y} \\ &= (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}.\end{aligned}$$

- Note that the solution $\hat{\boldsymbol{\beta}}$ minimizes

$$\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- Suppose Σ is a diagonal matrix of unequal error variances:

$$\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2).$$

- The GLS estimate of $\boldsymbol{\beta}$ minimizes

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{\sigma_i^2},$$

known as the Weighted LS (WLS).

- Errors are weighted proportional to $1/\sigma_i^2$: smaller weights for samples with larger variances.

- Suppose we have collected multiple obs at $\mathbf{x}_i$, $(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i})$. Let y_i denote the average of the n_i obs. Since

$$\sum_{j=1}^{n_i} (y_{ij} - \mathbf{x}_i^T \boldsymbol{\beta})^2 = n_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{n_i} (y_{ij} - y_i)^2,$$

it is enough to include one sample, $(\mathbf{x}_i, y_i)$, in the data. But $\text{Var}(y_i) = \sigma^2 / n_i$, not σ^2. So we should use WLS:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} n_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2.$$

- R command: `lm( ...., weights =` n_i`)`.

Justification via MLE

Model: $\mathbf{y} \sim \mathsf{N}_n\big(\mathbf{X}\boldsymbol{\beta}, \Sigma\big)$.

$$
\begin{aligned}
& \log p(\mathbf{y} \mid \boldsymbol{\beta}, \Sigma) \\
= \; & \log \left\{ \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \right\} \\
= \; & -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \text{ Constant}
\end{aligned}
$$

$$
\hat{\boldsymbol{\beta}}_{\mathsf{mle}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).
$$

GLS: Σ Unknown

How about this iterative approach?

- Start with some initial guess of Σ;
- Use Σ to estimate $\boldsymbol{\beta}$;
- Use residuals (since we've known $\boldsymbol{\beta}$) to estimate Σ;
- Iterate until convergence.

Not a bad idea. However, it won't work (actually no methods will work) if we know (or assume) nothing about Σ: too many parameters in Σ need to be estimated.

Usually, based on the application, we can assume the correlation structure, i.e., Σ, takes some particular form. Then, we can model Σ (now it does not involve too many parameters) and $\boldsymbol{\beta}$ simultaneously using likelihood based method. For example, for AR(1) times series,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots \\ \rho & 1 & \rho & \rho^2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & \cdots & 1 \end{pmatrix}.$$

Use the `nlme` package in R.

Testing for Lack of Fit

- How can we tell a model fits the data?
- If the model is correct then $\hat{\sigma}^2$ is an unbiased estimate of σ^2. So we can construct a test based on the ratio $\hat{\sigma}^2/\sigma^2$.
- Two cases: σ^2 known or unknown.

Lack of Fit: σ^2 Known

- H_0: no lack of fit; H_a: lack of fit.

- Test statistic
$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\mathsf{RSS}/(n-p)}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}.$$
Lack of fit means the error variance $\hat{\sigma}^2$ is large, i.e., large test statistic.

- Conclude that there is a lack of fit (i.e., Reject H_0), if
$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \geq \chi^2_{n-p}(1-\alpha).$$

Lack of Fit: σ^2 Unknown

- Get an estimate of σ^2 based on a very big/general model. And then derive the dist (under H_0) of the ratio $\hat{\sigma}^2_{\text{LinearModel}}/\hat{\sigma}^2_{\text{BigModel}}$. Basically we cast this problem as comparing two nested models.

- A general assumption (H_a): $y_i = f(\mathbf{x}_i) + \text{err}$.

- H_0: $y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \text{err}$.

- In order to operate this test, we need to have multiple obs at (at least) some $\mathbf{x}_i$'s,

$$(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i}), \quad i = 1 : m, \quad n = \sum_i n_i.$$

- $H_0 : y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta} + e_{ij}$ and e_{ij} iid $\sim \text{N}(0, \sigma^2)$. RSS_0 with df $=$ n-p.

- $H_a : y_{ij} = f(\mathbf{x}_i) + e_{ij}$ and e_{ij} iid $\sim \text{N}(0, \sigma^2)$ where f is any func.

$$\text{RSS}_a = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

with df $= n - m = \sum_i (n_i - 1)$.

- F-test

$$\frac{(\text{RSS}_0 - \text{RSS}_a)/(m-p)}{\text{RSS}_a/(n-m)} \sim F_{m-p, n-m}.$$