Statistical Inference for Independent Data

- Situation and model assumptions (scalar observations)
- Evaluation of model assumptions
- Approaches to inference
- What large sample theory says
- Conclusions for practice
- Multivariate observations
Situation and basic model assumptions

**Plan:** We consider the case of *scalar observations* first, then apply the insights gleaned to the case of *multivariate observations*.

**Recall:** We wrote a statistical model for pharmacokinetics of theophylline for a given subject as

\[ Y_j = f(t_j, U, \theta) + \epsilon_j, \quad j = 1, \ldots, n \]

\[ f(t, U, \theta) = \frac{k_aFD}{V(k_a - k_e)} \{ e^{-k_et} - e^{-k_at} \}, \quad \theta = (k_a, k_e, V)^T \]

- \( f(t, U, \theta) \) is the smooth function of \( t \) derived from the *deterministic* compartment model, \( U = \text{dose } D \text{ at } t = 0 \)
- \( \epsilon_j \) represents *deviation* that causes observations to not fall exactly on the smooth path \( f(t, U, \theta) \)
- \( Y = (Y_1, \ldots, Y_n)^T \) are observations taken at times \((t_1, \ldots, t_n)\) under conditions \( U \)
Situation and basic model assumptions

Conceptual representation:
Specific assumptions about $\epsilon_j$: Additive effects of measurement error, "biological fluctuations"

\[ \epsilon_j = \epsilon_{1j} + \epsilon_{2j} \]

- $\epsilon_{1j} \perp \epsilon_{2j} \mid U$
- $E(\epsilon_{1j} \mid U) = 0$, $\text{var}(\epsilon_{1j} \mid U) = \sigma_1^2$, $\epsilon_{1j} \perp \epsilon_{1j'} \mid U \Rightarrow \text{cov}(\epsilon_{1j}, \epsilon_{1j'} \mid U) = 0$
- $E(\epsilon_{2j} \mid U) = 0$, $\text{var}(\epsilon_{2j} \mid U) = \sigma_2^2$, $\text{cov}(\epsilon_{2j}, \epsilon_{2j'} \mid U) = \sigma_2^2 \exp\{-\phi(t_j - t_{j'})^2\}$
- Taken together $\Rightarrow \quad E(Y \mid U) = f(U, \theta)$, $\text{var}(Y \mid U) = \sigma_1^2 I_n + \sigma_2^2 \Gamma$ and $\text{var}(Y_j \mid U) = \sigma^2 = \sigma_1^2 + \sigma_2^2$

Common situation: The correlation among $Y_j$ is very small

- $t_j$ far apart in time – associations due to fluctuations have "died out"
- The effects of fluctuations are "dominated" by measurement error
Situation and basic model assumptions

**Approach:** *Reasonable approximations*

- $t_j$ far apart $\Rightarrow \epsilon_{2j} \perp \epsilon_{2j'} | U$ so that $\Gamma = I_n$ and thus

$$E(Y|U) = f(U, \theta), \quad \text{var}(Y|U) = \sigma^2 I_n$$

$$\sigma^2 = \sigma_1^2 + \sigma_2^2$$ is the *aggregate* variance due to measurement error and fluctuations at any $t_j$

- Measurement error *dominates* $\Rightarrow$ *eliminate* $\epsilon_{2j}$ from the model so

$$E(Y|U) = f(U, \theta), \quad \text{var}(Y|U) = \sigma^2 I_n$$

$$\sigma^2 = \sigma_1^2$$ is variance due to measurement error (the *only assumed source of variation*)

**Normality:** We also discussed assuming that $\epsilon_{1j}$ and $\epsilon_{2j}$ are *normally distributed* (conditional on $U$) so that $Y_j|U \sim \mathcal{N}\{f(t_j, U, \theta), \sigma^2\}$

- Do we *really need this*?
Situation and basic model assumptions

**Result:** For a *single series* of observations from a system at times $0 \leq t_1, < \ldots < t_n$, it is *common* to assume

$$E(Y_j|U) = f(t_j, U, \theta), \quad \text{var}(Y_j|U) = \sigma^2, \quad Y_j \text{ are } \perp \perp$$

- *In addition* may assume *normality*

- These are *assumptions* — they may not all be *correct* and should be considered carefully

- As we saw, *under these assumptions*, the *maximum likelihood estimator* for $\theta$ is the *ordinary least squares (OLS) estimator* minimizing

$$\sum_{j=1}^{n} \{Y_j - f(t_j, U, \theta)\}^2$$

- Thus, from a *statistical perspective*, the usual *inverse problem* implicitly makes such assumptions (that may or may not be true)
Basic model: for our discussion here, assume we are willing to believe

- Independence of $Y_j|U, j = 1, \ldots, n$
- $E(Y_j|U) = f(t_j, U, \theta)$
- And maybe more...

Questions:

- Is OLS the “best” thing to do?
- Can we do “better?”

Demonstration: We will consider the specific feature of variance of the observations as a device to illustrate the considerations involved
Evaluation of model assumptions

Recall: $\epsilon_j = Y_j - f(t_j, U, \theta)$

- **Routine assumption**: $\text{var}(Y_j|U) = \text{var}(\epsilon_j|U) = \sigma^2$, constant for all $j$

- OLS=MLE under *normality* $\Rightarrow$ this assumption *underlies* OLS

$$\sum_{j=1}^{n} \{y_j - f(t_j, U, \theta)\}^2$$

*More generally*, each $y_j$ receives “equal weight” in determining $\theta$

Why worry? *Subject-matter considerations* – it is *well-known* that for pharmacokinetic data

- *Assay error* is the *dominant source of variation*

- Error in determining concentrations is *greater* for *higher concentrations*
Evaluation of model assumptions

Can we check?

• If \( \text{var}(Y_j|U) \) really is a constant for all \( j \), the \( \epsilon_j \) would be of similar magnitude over the range of \( t_j \) or \( f(t_j, U, \theta) \)

Residuals: We cannot “see” \( \epsilon_j \), but we can get a sense of their values

\[
r_j = y_j - f(t_j, u, \hat{\theta}), \quad \hat{\theta} \text{ is the OLS estimate}
\]

• Plot \( r_j \) vs. \( t_j \) or \( f(t_j, u, \hat{\theta}) \) (the “predicted values”)

• If \( \text{var}(Y_j|U) = \sigma^2 \), expect to see a haphazard pattern about zero

• Common to plot the standardized residuals \( r_j/\hat{\sigma} \)

\[
\hat{\sigma}^2 = (n - p)^{-1} \sum_{j=1}^{n} \{y_j - f(t_j, u, \hat{\theta})\}^2
\]

• (Effect of estimating \( \theta \) rather than knowing it on plot… )
Evaluation of model assumptions

\[ r_j/\hat{\sigma} \text{ vs. } f(t_j, U, \hat{\theta}) : \]

Subject 12

Predicted Value

Residual

MA/ST 810 2009
Result: Magnitude of $r_j/\hat{\sigma}$ increases with $f(t_j, u, \hat{\theta})$

- **Assumption** $\text{var}(Y_j|U) = \sigma^2$ seems suspect

- A better assumption might be

$$\text{var}(Y_j|U) = \sigma^2 V\{f(t_j, U, \theta), \gamma\}, \quad V(\mu, \gamma) \uparrow \text{ in } \mu$$

- **Popular choice of** $V$: $V\{f(t_j, U, \theta), \gamma\} = f^{2\gamma}(t_j, U, \theta)$, $\gamma \geq 0$

- $\gamma = 1 \Rightarrow$ constant CV $\sigma$ – “Multiplicative error”

$$Y_j = f(t_j, U, \theta)(1 + \delta_j), \quad E(\delta_j|u) = 0, \quad \text{var}(\delta_j|U) = \sigma^2, \quad \epsilon_j = f(t_j, U, \theta)\delta_j$$

$\Rightarrow Y_j \sim \text{normal, lognormal, gamma} \ldots$

- $\gamma = 0.5 \Rightarrow$ “Poisson-like”

- Other $V$ models may also be used $\Rightarrow$ choice dictated by subject matter and empirical evidence
Implication: Assume instead

\[ E(Y_j | U) = f(t_j, U, \theta), \quad \text{var}(Y_j | U) = \sigma^2 f^2 \gamma(t_j, U, \theta) \]

or other suitable variance function

- \( \gamma \) may or may not be known

What about normality? Would expect residuals to show a symmetric pattern about zero

- Can be difficult to assess

- Does it matter? (coming up)
Reminder: *Even though* $\theta$ is of central interest, *must* get the *probability model correct*

- Here, $\psi = (\theta^T, \sigma^2, \gamma)^T$ are *all* the parameters

- If we are willing to believe a *particular distribution* (e.g., *normality*)
  $\Rightarrow$ *parametric model*

- If *not* $\Rightarrow$ *semiparametric model*

**Semiparametric model:** *Why* would we be unwilling to specify a distribution?

- *Outliers* – “extreme” observations occur more frequently than would expect under *normality*; *alternative distribution*?

- Features *may not correspond* to any known distribution

- *Fear of being wrong!*
Approaches to inference

**Parametric model:** Suppose we assume a model like

\[ Y_j | U \sim \mathcal{N}\{ f(t_j, U, \theta), \sigma^2 f^2 \gamma(t_j, U, \theta) \}, \quad Y_j \perp \perp \]

(1)

(the following applies equally to other variance models)

**Maximum likelihood estimation:** Under (1), the likelihood function for \( \psi = (\theta^T, \sigma^2, \gamma)^T \) is (suppress dependence on \( u \))

\[ L(\psi | y) \propto \prod_{j=1}^{n} \frac{1}{\sigma f \gamma(t_j, u, \theta)} \exp \left[ -\frac{\{y_j - f(t_j, u, \theta)\}^2}{2\sigma^2 f^2 \gamma(t_j, u, \theta)} \right] \]

- Maximizing \( L(\psi | y) \) in \( \psi \) clearly will not lead to the estimate of \( \theta \) gotten by minimizing \( \sum_{j=1}^{n} \{y_j - f(t_j, u, \theta)\}^2 \)

- This appears true even if we assumed we knew \( \gamma = 1 \) (or any other value \( \neq 0 \))
Approaches to inference

In particular: If \( \gamma = 0 \), we have \( \text{var}(Y_j|U) = \sigma^2 \), and recall

\[
\log L(\psi|y) = -n \log \sigma - \sum_{j=1}^{n} \frac{(y_j - f(t_j, u, \theta))^2}{2\sigma^2}
\]

⇒ maximizing \( L(\psi|y) \) is same as minimizing \( \sum_{j=1}^{n} (y_j - f(t_j, u, \theta))^2 \) or solving

\[
\frac{\partial}{\partial \theta} \log L(\psi|y) = \sum_{j=1}^{n} (y_j - f(t_j, u, \theta)) f_\theta(t_j, u, \theta) = 0 \quad (2)
\]

- \( \theta \) may be estimated “separately” from \( \sigma^2 \)

For \( \gamma \neq 0 \) (but known): \( \psi = (\theta^T, \sigma^2)^T \)

\[
\log L(\psi|y) = -n \log \sigma - \gamma \sum_{j=1}^{n} \log f(t_j, u, \theta) - \sum_{j=1}^{n} \frac{(y_j - f(t_j, u, \theta))^2}{2\sigma^2 f^{2\gamma}(t_j, u, \theta)}
\]

- Obviously must solve for \( \theta \) and \( \sigma^2 \) jointly
Aside: *Estimating equations*

- The MLE is found by *maximizing* an *objective function* (the likelihood function)
- When *derivatives exist*, maximizing an objective function is equivalent to *solving* a set of *equations*
- *However*, not *all estimators* with "good" properties are defined in this way ⇒ some are defined *directly* as the *solution to a set of equations* (especially for *semiparametric* models)
- *Moreover*, theoretical results (deriving *large sample approximate sampling distributions*) for both types may be based on casting estimators as solutions to equations
- Such equations are referred to as *estimating equations*
Approaches to inference

Estimating equations for MLE: $\gamma \neq 0$ known

$$\frac{\partial}{\partial \theta} \log L(\psi|y) = \sum_{j=1}^{n} \left\{ \frac{Y_j - f(t_j, U, \theta)}{f^2(\gamma(t_j, U, \theta))} \right\} f_\theta(t_j, U, \theta)$$

$$+ \sigma^2 \sum_{j=1}^{n} \left[ \left\{ \frac{Y_j - f(t_j, U, \theta)}{f^2(\gamma(t_j, U, \theta))} \right\}^2 - \sigma^2 f^2(\gamma(t_j, U, \theta)) \right] \left\{ \gamma \frac{f_\theta(t_j, U, \theta)}{f(t_j, U, \theta)} \right\} = 0 \quad (3)$$

$$\frac{\partial}{\partial \sigma^2} \log L(\psi|y) = \sum_{j=1}^{n} \left[ \left\{ \frac{Y_j - f(t_j, U, \theta)}{f^2(\gamma(t_j, U, \theta))} \right\}^2 - \sigma^2 f^2(\gamma(t_j, U, \theta)) \right] = 0 \quad (4)$$

- Must solve (3) and (4) **jointly** in $\theta$ and $\sigma^2$ — **can’t estimate $\theta$ separately**
- $\Rightarrow$ Although interest focuses on $\theta$, still must estimate **all parameters** in statistical model
- **Notice**: All of (2), (3), and (4) have **expectation zero under $\psi$** and the assumed statistical model and $\gamma$ value
Aside: If an estimating equation has expectation $= 0$ under an assumed statistical model, it is referred to as an unbiased estimating equation.

- Under “nice” conditions, estimators solving unbiased estimating equations are consistent.
- If an estimating equation is biased, it may lead to inconsistent estimator.
- More coming up...
Approaches to inference

**Alternative perspective:** When $\gamma = 0$ [so $\text{var}(Y_j|U) = \sigma^2$], OLS may also be motivated as “*minimize distance between data and model*”

$$
\sum_{j=1}^{n} \left\{ y_j - f(t_j, u, \theta) \right\}^2 \Rightarrow \sum_{j=1}^{n} \left\{ y_j - f(t_j, u, \theta) \right\} f_{\theta}(t_j, u, \theta) = 0
$$

- $\text{var}(Y_j|U) = \sigma^2$ constant $\Rightarrow Y_j$ of “equal quality” $\Rightarrow$ “equal weight”

**Suggestion:** Suppose $\text{var}(Y_j|U) = \sigma^2/w_j$ for known constants $w_j$

- **Under normality assumption** – MLE for $\theta$ minimizes

$$
\sum_{j=1}^{n} w_j \left\{ y_j - f(t_j, u, \theta) \right\}^2 \Rightarrow \sum_{j=1}^{n} w_j \left\{ y_j - f(t_j, u, \theta) \right\} f_{\theta}(t_j, u, \theta) = 0 \quad (5)
$$

- **Weighted least squares (WLS):** Even without normality $\Rightarrow$ “*minimize distance between data and model taking into account unequal quality*” $\Rightarrow$ “unequal weight”

- I.e., $\text{var}(Y_j|U)$ small when $w_j$ large $\Rightarrow$ “*higher quality (precision)*”
**Approaches to inference**

**Generalized least squares:** \( \text{var}(Y_j|U) = \sigma^2 f^{2\gamma}(t_j, U, \theta) \Rightarrow w_j = f^{-2\gamma}(t_j, U, \theta) \)

- \( w_j \) *not* constants (depend on \( \theta \)), but (5) suggests solving

\[
\sum_{j=1}^{n} f^{-2\gamma}(t_j, u, \theta) \{y_j - f(t_j, u, \theta)\} f(\theta)(t_j, u, \theta) = 0 \quad (6)
\]

\( \Rightarrow \) “Generalized least squares (GLS) estimator”

- Nice feature – *does not involve* \( \sigma^2 \)

- *Important*: Does *not necessarily* correspond to max/minimizing an *objective function*

- E.g., (6) *does not* result from minimizing

\[
\sum_{j=1}^{n} f^{-2\gamma}(t_j, u, \theta) \{y_j - f(t_j, u, \theta)\}^2
\]

\( \Rightarrow \) minimizing this leads to a *inconsistent estimator*
Approaches to inference

**GLS facts:**

- Solving (6) is *straightforward*: “*Iteratively ReWeighted Least Squares (IRWLS)*”

- Does not necessarily come from any particular *distributional* assumption, *only* from assumptions on \( E(Y_j|U) \), \( \text{var}(Y_j|U) \)

- Thus, is a natural choice for *semiparametric models*

- *Actually*, for \( \gamma = 1 \), is MLE assuming \( Y_j|U \) have a *gamma* distribution; for \( \gamma = 0.5 \), is MLE assuming *Poisson* distribution

- Can estimate \( \sigma^2 \) at the end by

\[
\hat{\sigma}^2 = (n - p)^{-1} \sum_{j=1}^{n} f^{-2\gamma}(t_j, u, \hat{\theta}) \{y_j - f(t_j, u, \hat{\theta})\}^2
\]

where \( \hat{\theta} \) is the GLS estimate ⇒ see (4)
Approaches to inference

Comparing GLS and normal MLE:

- Recall from (3) the MLE for $\theta$ solves [jointly with (4)]

$$
\sum_{j=1}^{n} \frac{\{y_j - f(t_j, u, \theta)\}}{f^{2\gamma}(t_j, u, \theta)} f_\theta(t_j, \theta)
$$

$$
+ \sigma^2 \sum_{j=1}^{n} \left[ \frac{\{y_j - f(t_j, u, \theta)\}^2 - \sigma^2 f^{2\gamma}(t_j, u, \theta)}{\sigma^2 f^{2\gamma}(t_j, u, \theta)} \right] \left\{ \gamma \frac{f_\theta(t_j, u, \theta)}{f(t_j, u, \theta)} \right\} = 0
$$

- GLS for $\theta$ solves

$$
\sum_{j=1}^{n} \frac{\{y_j - f(t_j, u, \theta)\}}{f^{2\gamma}(t_j, u, \theta)} f_\theta(t_j, u, \theta) = 0
$$

- *First term* of MLE equation is *the same* as the GLS equation!

- This term is *linear* in the $y_j$

- The *second* MLE term is *quadratic* in $y_j$

- These features have implications for properties...
Approaches to inference

What about $\gamma$? Often, may be unwilling to specify a fixed numerical value for $\gamma$

- For many assays, $\gamma \approx 0.7–0.9$ seems more appropriate than “standard” values like 0.5, 1.0 – no well-known distribution with these values

- So may not be willing to adopt distributional model like the gamma probability distribution for which $\gamma$ is fixed

- For this and other variance models, may just have no idea; e.g.,

$$\text{var}(Y_j | U) = \sigma^2 V \{ f(t_j, U, \theta), \gamma \} = \sigma^2 \{ \gamma_1 + f^2 \gamma_2(t_j, U, \theta) \}$$

- It is possible to derive estimating equation for $\gamma$ to be solved jointly with $\sigma^2$, $\theta$

- If willing to assume normality, can maximize likelihood jointly
What large sample theory says

**Issues:** Assume we’ve at least got \( f(t, u, \theta) \) correct

1. If variance is really *nonconstant*, but use OLS estimator *anyway*, what are the consequences (usual *inverse problem*)?

2. If we *acknowledge* and *model* nonconstant variance, how to choose between GLS and MLE assuming we’ve modeled variance *correctly*? In particular, what if we use MLE assuming *normality* and we’re *wrong*?

3. What if we’ve modeled variance *incorrectly*?

**To address these questions:**

- *Cannot* get *exact* (finite \( n \)) results for *sampling properties*
- \( \Rightarrow \) Derive large-sample approximate *sampling distributions*
- *Compare* on the basis of *consistency*, ARE
What large sample theory says

Suppose: \( \text{var}(Y_j|U) = \sigma^2 V \{ f(t_j, U, \theta), \gamma \} \)

1. Results for OLS: The model says that if \( \psi = (\theta^T, \sigma^2)^T \) is the parameter value, \( E_{\psi}(Y_j|U) = f(t_j, U, \theta) \)

   - **OLS estimating equation** \( \sum_{j=1}^{n} \{ Y_j - f(t_j, U, \theta) \} f_\theta(t_j, U, \theta) = 0 \)

   - \( E_{\psi} \left[ \{ Y_j - f(t_j, U, \theta) \} f_\theta(t_j, U, \theta) | U \right] = 0 \) under the model REGARDLESS of the true form of \( \text{var}(Y_j|U) \)

   - \( \Rightarrow \) **Unbiased estimating equation**

   - Thus, \( \hat{\theta}_{OLS} \xrightarrow{p} \theta_0 \) even if variance is nonconstant

So why bother? **Sampling distribution**

(i) Calculation of *valid* approximate standard errors, confidence intervals
   (accurate assessment of uncertainty)

(ii) **Relative inefficiency**
What large sample theory says

Large-$n$ approximate sampling distribution of $\hat{\theta}_{OLS}$:

- Define (suppress $U$)
  \[
  F_\theta(\theta) = \begin{pmatrix}
  f_\theta^T(t_1, U, \theta) \\
  \vdots \\
  f_\theta^T(t_n, U, \theta)
  \end{pmatrix},
  \]
  \[
  W^{-1}(\theta, \gamma) = \text{diag}[V\{f(t_1, U, \theta), \gamma\}, \ldots, V\{f(t_n, U, \theta), \gamma\}]
  \]

- If IN TRUTH, $\text{var}(Y_j|U) = \sigma^2$ constant, then (conditional on $U$)
  \[
  \hat{\theta}_{OLS} \sim \mathcal{N}_p\{\theta_0, \sigma_0^2 \Sigma_{OLS}(\theta_0)\} \tag{7}
  \]
  \[
  \Sigma_{OLS}(\theta_0) = \{F_\theta^T(\theta_0)F_\theta(\theta_0)\}^{-1}
  \]

- Standard errors based on (7) $\Rightarrow$ substitute $\hat{\theta}_{OLS}$ and
  \[
  \hat{\sigma}_{OLS}^2 = (n - p)^{-1}\sum_{j=1}^{n}\{y_j - f(t_j, u, \hat{\theta}_{OLS})\}^2
  \]
What large sample theory says

Large-$n$ approximate sampling distribution of $\hat{\theta}_{OLS}$:

- **HOWEVER**, if **IN TRUTH**, $\text{var}(Y_j|U) = \sigma^2 V\{f(t_j, U, \theta), \gamma\}$, then

$$\hat{\theta}_{OLS} \sim \mathcal{N}_{p}\{\theta_0, \sigma_0^2 \Sigma_W(\theta_0, \gamma)\}$$  \hspace{1cm} (8)

$$\Sigma_W(\theta_0, \gamma) = \{F_T^\theta(\theta_0)F(\theta_0)\}^{-1}\{F_T^\theta(\theta_0)W^{-1}(\theta_0, \gamma)F(\theta_0)\}\{F_T^\theta(\theta_0)F(\theta_0)\}^{-1}$$

- Clearly **different from** $\Sigma_{OLS}(\theta_0)$

- That is, the properties of $\hat{\theta}_{OLS}$ are **different** depending on whether or not the **constant variance** assumption is really **correct**

- Thus, under these conditions, (7) is **NOT** a valid approximation to the sampling distribution of $\hat{\theta}_{OLS}$
What large sample theory says

(i) Calculation of accurate assessment of uncertainty:

- *Standard software* assumes that if we use OLS, it’s because we *think* we have *constant variance* [so uses (7)]

\[ \Rightarrow \]

- If we estimate \( \theta \) by \( \hat{\theta}_{OLS} \) and use (7) for standard errors, etc, and *IN TRUTH* constant variance *does not hold*, assessments of uncertainty will be *flawed* (and usually *optimistic*)

- To obtain *correct* assessment, must use (8), which depends on *true variance function*!

- There is a *way* of getting *valid* standard errors using (8), but we *still* have issue (ii) . . .

(ii) Efficiency considerations: Compare OLS with GLS and MLE . . .
What large sample theory says

Large-\(n\) approximate sampling distribution of \(\hat{\theta}_{GLS}\):

- If \(E_\psi(Y_j|U) = f(t_j, U, \theta)\), the \textit{GLS estimating equation}

\[
\sum_{j=1}^{n} V^{-1}\{f(t_j, U, \theta), \gamma\}\{Y_j - f(t_j, U, \theta)\}f_\theta(t_j, U, \theta) = 0
\]

is \textit{unbiased REGARDLESS} of whether we have \(\text{var}(Y_j|U)\) correct

- \(\Rightarrow \hat{\theta}_{GLS}\) is \textit{consistent} even if we are wrong about \(\text{var}(Y_j|U)\)

- If we are \textit{correct} that \(\text{var}_\psi(Y_j|U) = \sigma^2 V\{f(t_j, U, \theta), \gamma\}\), then (conditionally on \(U\))

\[
\hat{\theta}_{GLS} \sim \mathcal{N}_p\{\theta_0, \sigma_0^2 \Sigma_{WLS}(\theta_0, \gamma)\} \quad (9)
\]

\[
\Sigma_{WLS}(\theta_0, \gamma) = \{F_\theta^T(\theta_0) W(\theta_0, \gamma) F_\theta(\theta_0)\}^{-1}
\]
What large sample theory says

Large-\(n\) approximate sampling distribution of \(\hat{\theta}_{GLS}\):

- **Easy to show**: Sampling variance in (8) \(\geq\) sampling variance in (9)\n  \(\Rightarrow\) ARE of \(\hat{\theta}_{GLS}\) to \(\hat{\theta}_{OLS}\) \(\geq 1\) when \(\text{var}_\psi(Y_j|U)\) is *correctly specified*

- So \(\hat{\theta}_{GLS}\) is *more precise* than \(\hat{\theta}_{OLS}\) in general

- If we estimate \(\theta\) using \(\hat{\theta}_{GLS}\), we may obtain *standard errors* using (9) by substituting estimates for \(\sigma^2_0\) and \(\theta_0\) (and \(\gamma\) if it is also estimated)

- Of course, this assumes we have the variance model *correctly specified*
What large sample theory says

(ii) Efficiency considerations: Result – if we model nonconstant variance and do a good job, \( \hat{\theta}_{GLS} \) is more precise than \( \hat{\theta}_{OLS} \) (for \( n \) “large”)

- Using \( \hat{\theta}_{OLS} \) can result in weaker conclusions
- This continues to be true even if we also have to estimate \( \gamma \) (large \( n \))
- The same comparison holds with normal MLE...
2. **GLS vs. MLE:** Suppose our choice \( \text{var}(Y_j | U) = \sigma^2 V \{ f(t_j, U, \theta), \gamma \} \) is correct AND \( Y_j | U \) really is normally distributed

- \( \hat{\theta}_{MLE} \) solves
  \[
  \sum_{j=1}^{n} \frac{Y_j - f(t_j, U, \theta)}{V \{ f(t_j, U, \theta), \gamma \}} f_\theta(t_j, U, \theta) \]

  \[
  + \frac{\sigma^2}{2} \sum_{j=1}^{n} \left[ \frac{\{Y_j - f(t_j, U, \theta)\}^2 - \sigma^2 V \{ f(t_j, U, \theta), \gamma \}}{\sigma^2 V \{ f(t_j, U, \theta), \gamma \}} \right] \frac{V_\theta \{ f(t_j, U, \theta), \gamma \}}{V \{ f(t_j, U, \theta), \gamma \}} = 0
  \]

- While \( \hat{\theta}_{GLS} \) solves
  \[
  \sum_{j=1}^{n} \frac{Y_j - f(t_j, U, \theta)}{V \{ f(t_j, U, \theta), \gamma \}} f_\theta(t_j, U, \theta) = 0
  \]

- **Both** estimators are \textit{consistent under these conditions}
What large sample theory says

2. GLS vs. MLE:

• For large $n$ and some positive definite $\Lambda(\theta_0, \gamma)$

$$\hat{\theta}_{GLS} \sim \mathcal{N}_p\{\theta_0, \sigma_0^2 \Sigma_{WLS}(\theta_0, \gamma)\}$$

$$\Sigma_{WLS}(\theta_0, \gamma) = \{F_{\theta}^T(\theta_0)W(\theta_0, \gamma)F_{\theta}(\theta_0)\}^{-1}$$

$$\hat{\theta}_{MLE} \sim \mathcal{N}_p\{\theta_0, \sigma_0^2 \Sigma_{MLE}(\theta_0, \gamma, \sigma_0^2)\}$$

$$\Sigma_{MLE}(\theta_0, \gamma, \sigma_0^2) = \{\Sigma_{WLS}^{-1}(\theta_0, \gamma) + 2\sigma_0^2 \Lambda(\theta_0, \gamma)\}^{-1}$$

• $\Rightarrow$ ARE of $\hat{\theta}_{MLE}$ to $\hat{\theta}_{GLS}$ is $\geq 1$
What large sample theory says

\textbf{Result:} If $Y_j|U$ are \textit{exactly} normally distributed \textit{AND} 
\[ \text{var}(Y_j|U) = \sigma^2 V \{ f(t_j, U, \theta, \gamma) \} \] is \textit{correct}, $\hat{\theta}_{MLE}$ is \textit{more precise}

However:

\begin{itemize}
  \item If $Y_j|U$ are \textit{NOT} normally distributed, the advantage is \textit{lost} (the form of 
  $\Sigma_{MLE}(\theta_0, \gamma, \sigma^2_0)$ changes)
  \item \textit{In particular}, if there are \textit{OUTLIERS}, the \textit{efficiency loss} using $\hat{\theta}_{MLE}$ 
  relative to $\hat{\theta}_{GLS}$ can be \textit{substantial}
  \item \textit{REASON}: Because of \textit{quadratic dependence} of MLE estimating equation 
  on $Y_j$, properties of $\hat{\theta}_{MLE}$ depend on \textit{MORE} properties of the \textit{true} 
  distribution of $Y_j|U$ than do those of $\hat{\theta}_{GLS}$ (up to \textit{fourth moments})
  \item Because of only \textit{linear dependence} on $Y_j$, properties of $\hat{\theta}_{GLS}$ depend 
  \textit{ONLY} on the assumed $E(Y_j|U)$ and $\text{var}(Y_j|U)$, and \textit{nothing more} $\Rightarrow$ 
  \textit{properties are the same whether $Y_j|U$ is normal or not} (the form of 
  $\Sigma_{GLS}(\theta_0, \gamma)$ is the \textit{same} regardless)
\end{itemize}
What large sample theory says

3. What if variance is incorrectly specified? More bad news

- \( \hat{\theta}_{MLE} \) solves
  \[
  \sum_{j=1}^{n} \frac{Y_j - f(t_j, U, \theta)}{V\{f(t_j, U, \theta), \gamma\}} f_\theta(t_j, U, \theta)
  \]
  \[+
  \frac{\sigma^2}{2} \sum_{j=1}^{n} \left[ \frac{(Y_j - f(t_j, U, \theta))^2}{\sigma^2 V\{f(t_j, U, \theta), \gamma\}} - \frac{\sigma^2 V\{f(t_j, U, \theta), \gamma\}}{\sigma^2 V\{f(t_j, U, \theta), \gamma\}} \right] \frac{V\{f(t_j, U, \theta), \gamma\}}{V\{f(t_j, U, \theta), \gamma\}} = 0
  \]

- If the chosen model \( \sigma^2 V\{f(t_j, U, \theta), \gamma\} \) for \( \text{var}(Y_j|U) \) is incorrect, the MLE estimating equation is no longer unbiased

- \( \Rightarrow \hat{\theta}_{MLE} \) will be inconsistent

- \( \hat{\theta}_{GLS} \) solves
  \[
  \sum_{j=1}^{n} \frac{Y_j - f(t_j, U, \theta)}{V\{f(t_j, U, \theta), \gamma\}} f_\theta(t_j, U, \theta) = 0
  \]

- \( \Rightarrow \) Estimating equation is still unbiased, \( \hat{\theta}_{GLS} \) will be consistent
What large sample theory says

**Result:** GLS is "robust to" misspecification of the variance structure, MLE is *NOT*

**Special situation:** "High-quality data"

- As in the theophylline subject 12 data
- \( \text{var}(Y_j|U) \) is *small relative to the range* of \( f(t_j, U, \theta) \) values studied
  \[ \Rightarrow \text{if var}(Y_j|U) = \sigma^2 V\{f(t_j, U, \theta), \gamma\}, \text{represent by } \sigma \text{ "small" (i.e., } \sigma \to 0) \]
- Under these conditions, \( \hat{\theta}_{GLS} \) and \( \hat{\theta}_{MLE} \) are *asymptotically equivalent*
- \( \Rightarrow \) In practice (finite \( n \)), *very similar* inferences
- Normal MLE used heavily in pharmacokinetics, toxicokinetics
- "Extended least squares"
Conclusions for practice

Moral:

- OLS is *OFTEN NOT* the preferred approach (*inverse problem*)
- Minimizing the OLS objective function will lead to *inefficient inferences* and potentially *erroneous conclusions* when the variance is *not constant* . . .
- . . . which is the case in practice *often* (especially for *biological systems*)
- MLE based on assumption of *normality* is *sensitive* to *violation of assumptions*
- GLS is “*safer*”
- GLS *does not necessarily* correspond to min/maximizing an *objective function*, although it can be implemented this way
Conclusions for practice

Remarks:

- It may in fact be shown that the GLS estimator is “asymptotically optimal” among all estimators for $\theta$ in the semiparametric model where only $E(Y_j|U)$ and var$(Y_j|U)$ are specified.

- Note that we must estimate all components of $\psi$, including those not of central interest – efficiency of estimation of $\theta$ can depend on that of estimators for “nuisance parameters” like $\gamma$ when it is unknown.

- Methods for estimating variance parameters like $\gamma$ are available; involve solving additional estimating equations.
Remarks: Alternative approach to nonconstant variance – transformation

- Place both observations and model on a transformed scale where variance is thought to be constant

- E.g., for suspected constant coefficient of variation

\[
\log Y_j = \log \{f(t_j, U, \theta)\} + \epsilon_j^*, \quad \text{var}(\epsilon_j^*|U) = \sigma^2; \quad (10)
\]

maybe assume normality on this scale, too

- ⇒ Use OLS on this scale to estimate \( \theta \)

- Is approximately equivalent to assuming \( \text{var}(Y_j|U) = \sigma^2 f^2(t_j, U, \theta) \)

- Same issues: Wrong transformation is like wrong variance model
Conclusions for practice

Remarks:

• All of this has been *predicated* on the assumption of *independence*.

• If *serial correlation* (i.e., correlation over *time*) is *nonnegligible*, must take into account ⇒ considerations of *time series* modeling and analysis enter the picture...

• ...can examine *residuals* over time and *model* correlation under *stationarity assumptions* (beyond scope of our discussion here).

• If assumption of *negligible correlation* is *inappropriate*, fancier methods needed.
Multivariate observations

Remarks: Model selection

- It is important to recognize that there are two models: the structural mathematical model and the statistical model in which it is embedded in this framework.

- Routine objectives as part of an inverse problem: Evaluation of suitability of the mathematical model, sensitivity analysis.

- If the mathematical model is embedded in an incorrect statistical model, conclusions drawn may be erroneous.

Extension: These results extend to multivariate observations, where $Y_j$ and $f(t_j, U, \theta)$ are $(k \times 1)$ vectors (up next).
Multivariate observations

**General setting:** \( Y = (Y_1^T, \ldots, Y_n^T) \) at times \((t_1, \ldots, t_n)\), where

\[
Y_j = \left( Y_j^{(1)}, \ldots, Y_j^{(k)} \right)^T \perp \perp \text{across } j
\]
is observation at time \(t_j\) on some function of states of math model \(x(t)\)

- \( f(t, U, \theta) \) found via **observation matrix** \(\mathcal{O}\)

\[
f(t, U, \theta) = \mathcal{O}x(t, U, \theta) = \begin{pmatrix}
f^{(1)}(t, U, \theta) \\
\vdots \\
f^{(k)}(t, U, \theta)
\end{pmatrix}
\]

- **Usual simplified statistical model** assuming all intra-subject correlations **negligible** and **constant variances**

\[
Y_j|U \sim \mathcal{N}_k\{f(t_j, U, \theta), \mathcal{V}(\sigma^2)\}, \quad j = 1, \ldots, n
\]

\[
\mathcal{V}(\sigma^2) = \text{diag}(\sigma^{(1)^2}, \ldots, \sigma^{(k)^2}) \quad \psi = (\theta^T, \sigma^{(1)^2}, \ldots, \sigma^{(k)^2})^T
\]
MLE under normality: Assuming this *parametric model*, the MLE for $\theta$ is found by maximizing in $\psi$

$$
\log L(\psi|y) = -n \log |\mathcal{V}| - \sum_{j=1}^{n} \{y_j - f(t_j, u, \theta)\}^T \mathcal{V}^{-1}(\sigma^2) \{y_j - f(t_j, u, \theta)\}
$$

$$
= -(n \log \sigma(1)^2 + \cdots + n \log \sigma(k)^2) - \left[ \frac{1}{\sigma(1)^2} \sum_{j=1}^{n} \{y_j^{(1)} - f^{(1)}(t_j, u, \theta)\}^2 + \cdots + \frac{1}{\sigma(k)^2} \sum_{j=1}^{n} \{y_j^{(k)} - f^{(k)}(t_j, u, \theta)\}^2 \right]
$$

- Can *no longer separate* estimation of $\theta$ from that of the variances $\sigma(1)^2, \ldots, \sigma(k)^2$
Multivariate observations

Result: Must solve jointly

\[ \hat{\theta}_{GLS} = \arg \min_{\theta} \left[ \frac{1}{\sigma^{(1)}^2} \sum_{j=1}^{n} \{y_j^{(1)} - f^{(1)}(t_j, u, \theta)\}^2 + \cdots + \frac{1}{\sigma^{(k)}^2} \sum_{j=1}^{n} \{y_j^{(k)} - f^{(k)}(t_j, u, \theta)\}^2 \right] \]

\[ \sigma^{(\ell)^2} = n^{-1} \sum_{j=1}^{n} \{y_j^{(\ell)} - f^{(\ell)}(t_j, u, \theta)\}^2, \quad \ell = 1, \ldots, k \]

- **Practical interpretation**: Each component of \( y_j \) is weighted in accordance with its assumed (constant) variance

- A version of “weighted least squares” with “estimated weights” (estimated constant variances)

- A version of generalized least squares

- Differentiation yields an unbiased estimating equation for \( \theta \Rightarrow \hat{\theta}_{GLS} \) will be consistent for the true value \( \theta_0 \)
Multivariate observations

Contrast with: “OLS”

\[ \hat{\theta}_{OLS} = \arg \min_{\theta} \sum_{j=1}^{n} \{y_j - f(t_j, u, \theta)\}^T \{y_j - f(t_j, u, \theta)\} \]

\[ = \arg \min_{\theta} \sum_{\ell=1}^{k} \sum_{j=1}^{n} \{y_{j}^{(\ell)} - f^{(\ell)}(t_j, u, \theta)\}^2 \]

• “Equal weighting”: Corresponds to taking \( V(\sigma^2) = \sigma^2 I_k \)

• Also yields an unbiased estimating equation so \( \hat{\theta}_{OLS} \) will also be consistent

• BUT by analogy to results for the scalar case, \( \hat{\theta}_{OLS} \) will be inefficient relative to \( \hat{\theta}_{GLS} \) if the components \( Y_{j}^{(\ell)} \) of \( Y_j \) really do have different (constant) variances given \( U \)
Multivariate observations

Large-\(n\) approximate sampling distribution of \(\hat{\theta}_{GLS}\):

- Define
  \[
  D_{\theta_j}(\theta) = \frac{\partial}{\partial \theta} f(t_j, U, \theta) = \begin{pmatrix}
  f^{(1)}_\theta(T(t_j, U, \theta)) \\
  \vdots \\
  f^{(k)}_\theta(T(t_j, U, \theta))
  \end{pmatrix}, \quad (k \times p),
  \]

  the matrix of partial derivatives of the \(k\) elements of \(f(t_j, U, \theta)\) with respect to the elements of \(\theta\) \((p \times 1)\)

- If the assumption on variances in model (11) is correct; i.e., \(\text{var}(Y_j|U) = \mathcal{V}(\bar{\sigma}^2)\), then with \(\bar{\sigma}^2 = (\sigma^{(1)}^2, \ldots, \sigma^{(k)}^2)^T\),

  \[
  \hat{\theta}_{GLS} \sim \mathcal{N}_p\{\theta_0, \Sigma_{GLS}(\theta_0, \bar{\sigma}_0^2)\}
  \]

  \[
  \Sigma_{GLS}(\theta_0, \bar{\sigma}_0^2) = \left\{ \sum_{j=1}^{n} D_{\theta_j}(\theta_0)^T \mathcal{V}^{-1}(\bar{\sigma}_0^2) D_{\theta_j}(\theta_0) \right\}^{-1}
  \]

- Can obtain standard errors by substituting estimates as usual
Multivariate observations

**More complex models:** It may be the case that some (or all) components of \( Y_j \) have *nonconstant variance*; i.e., we may wish instead to assume

\[
E(Y_j^{(\ell)} | U) = f^{(\ell)}(t_j, U, \theta), \quad \text{var}(Y_j^{(\ell)} | U) = \sigma^{(\ell)2} \{ f^{(\ell)}(t_j, U, \theta) \}^{2\gamma^{(\ell)}}
\]

for each \( \ell = 1, \ldots, k \), so that

\[
E(Y_j|U) = f(t_j, U, \theta), \quad \text{var}(Y_j|U) = \mathcal{V}_j(\theta, \overline{\sigma}^2, \overline{\gamma})
\]

\[
\mathcal{V}_j(\theta, \overline{\sigma}^2, \overline{\gamma}) = \text{diag}[\sigma^{(1)2} \{ f^{(1)}(t_j, U, \theta) \}^{2\gamma^{(1)}}, \ldots, \sigma^{(k)2} \{ f^{(k)}(t_j, U, \theta) \}^{2\gamma^{(k)}}] \]

\[
\overline{\sigma}^2 = (\sigma^{(1)2}, \ldots, \sigma^{(k)2})^T, \quad \overline{\gamma} = (\gamma^{(1)}, \ldots, \gamma^{(k)})^T
\]

\[
\psi = (\theta^T, \overline{\sigma}^2, \overline{\gamma})^T
\]
GLS estimator: Solves the (unbiased) estimating equations

\[
\sum_{j=1}^{n} D_{\theta_j}^{T}(\theta) V_j^{-1}(\theta, \sigma^2, \gamma) \{Y_j - f(t_j, U, \theta)\} = 0
\]

jointly with equations for \( \sigma^2 \) (and \( \gamma \) if unknown)

• For \( \gamma \) known, solve with

\[
\sigma^{(\ell)^2} = (n-p)^{-1} \sum_{j=1}^{n} \{f^{(\ell)}(t_j, U, \theta)\}^{-2\gamma^{(\ell)}} \{Y_j^{(\ell)} - f_j^{(\ell)}(t_j, U, \theta)\}^2, \quad \ell = 1, \ldots, k
\]
Multivariate observations

Large-\( n \) approximate sampling distribution of \( \hat{\theta}_{GLS} \):

- If the \textit{variance models} for each component are all \textit{correctly specified}
  \[
  \hat{\theta}_{GLS} \sim \mathcal{N}_p\{\theta_0, \Sigma_{GLS}(\theta_0, \sigma^2, \gamma)\}
  \]
  \[
  \Sigma_{GLS}(\theta_0, \sigma^2, \gamma) = \left\{ \sum_{j=1}^{n} D_{\theta j}^T(\theta_0) V_j^{-1}(\theta_0, \sigma^2, \gamma) D_{\theta j}(\theta_0) \right\}^{-1}
  \]

- Can obtain \textit{standard errors} by substituting estimates as usual
Multivariate observations

**MLE vs. GLS:** If we furthermore were to assume *normality*, an alternative is to estimate $\theta$ and $\sigma^2$ (and $\gamma$ if unknown) by *maximum likelihood*

- *Same issues* as in the scalar case
- *Quadratic vs. linear* estimating equations
- MLE is *more precise asymptotically* if model and normality *exactly correct*... 
- ...but is sensitive to *violations of assumptions* (can be inefficient, inconsistent)
- Again, GLS is “*safer*”