Principles of Statistical Inference

- Recap of statistical models
- Statistical inference (frequentist)
- Parametric vs. semiparametric models and inference
- Large sample approximation
Recap of statistical models

Theophylline example: Recap

- **Mathematical model**

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

\[U = D = \text{oral dose at } t = 0\]

- **Statistical model with intra-subject correlation**

\[
Y|U \sim \mathcal{N}_n\{f(U, \theta), \sigma_1^2I_n + \sigma_2^2\Gamma\}, \quad \psi = (\theta^T, \sigma_1^2, \sigma_2, \phi)^T
\] (1)

- Simplified statistical model assuming *negligible* intra-subject correlation

\[
Y|U \sim \mathcal{N}_n\{f(U, \theta), \sigma^2I_n\}, \quad \psi = (\theta^T, \sigma^2)^T
\] (2)
Recap of statistical models

**HIV example:** Recap

- \( Y = (Y_1^T, \ldots, Y_n^T) \) at times \((t_1, \ldots, t_n)\), where
  \[
  Y_j = (Y_j^{(1)}, Y_j^{(2)})^T \quad \text{bivariate}
  \]
  is observation at time \( t_j \) on
  \[
  \{T_1(t_j) + T_1^*(t_j), V_I(t_j) + V_{NI}(t_j)\}^T = \text{(total CD4, viral load)}
  \]

- **Mathematical model:** Not repeated here for brevity; 7 states
  \[
  x(t) = \{T_1(t), T_2(t), T_1^*(t), T_2^*(t), V_I(t), V_{NI}(t), E(t)\}^T
  \]
  and observed continuous time **ARV input** \( U = U(t) \)
  \[
  f(t, U, \theta) = \mathcal{O}x(t, U, \theta) = \begin{pmatrix} f^{(1)}(t, U, \theta) \\ f^{(2)}(t, U, \theta) \end{pmatrix}
  \]
Recap of statistical models

HIV example: Recap

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

\[
Y_j | U \sim \mathcal{N}_2 \{f(t, U, \theta), \mathcal{V}\}, \quad j = 1, \ldots, n
\]  

\[
\mathcal{V} = \begin{pmatrix}
\sigma^{(1)^2} & 0 \\
0 & \sigma^{(2)^2}
\end{pmatrix}, \quad \psi = (\theta^T, \sigma^{(1)^2}, \sigma^{(2)^2})^T
\]

- **More complicated** statistical models of course possible; (3) is often *tacitly assumed* in inverse problems without comment regarding sources of variation
Recall: A statistical model like (1), (2), or (3) describes the assumed family of all possible probability distributions for random vector $Y$ representing the data generating mechanism for observations we might see at $t_1, \ldots, t_n$ under conditions $U$

- Indexed by a parameter vector $\psi$

- We are interested in the component $\theta$ of $\psi$ that governs the mathematical model, but the other components are needed to describe fully the statistical model, so we are stuck dealing with all of $\psi$

- The assumed statistical model may or may not be correct
Statistical inference

For definiteness: We discuss basic principles of *statistical inference* in the context of *statistical models* of the form

\[ Y_j = f(t_j, U, \theta) + \epsilon_j, \]

where \( Y_j \) is *scalar* (generalize to *multivariate* \( Y_j \) later...)

- \((Y_1, \ldots, Y_n)\) are *independent* (conditional on \( U \)) with

\[
Y|U = (Y_1, \ldots, Y_n)^T \sim \mathcal{N}_n \{f(U, \theta), \sigma^2 I_n\}, \quad \psi = (\theta^T, \sigma^2)^T,
\]

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

- I.e., \( Y_j, j = 1, \ldots, n, \) may be regarded as representing the result of a potential “draw” from the *normal probability distribution* describing how such observations would “turn out” at \( t_j \) (given phenomena like measurement error and “biological fluctuation”) under conditions \( U \)
Statistical inference

Conceptually:

- The statistical model is a *formal representation* of the “population” of *all possible realizations* of \((Y_1, \ldots, Y_n)\) we would ever see under \(U\).

- When we collect data, we observe a *sample* from the *population*; i.e., we get to see a *single realization* of \((Y_1, \ldots, Y_n), (y_1, \ldots, y_n)\) (and a realization \(u\) of \(U\), say).

**Objective, restated:** What can we learn about the “true value” of \(\psi\) (which determines the nature of the *population*) from a *sample*?

- Do not get to observe the entire population (or else we’d know \(\psi\)).

- How *uncertain* will we be?

- What if we’re *wrong* about the statistical model (later)?

**Statistical inference (loosely speaking):** Making statements about a *population* of interest on the basis of only a *sample* from the population.
**Statistic:** Any *function* of a random variable(s)

**Parameter (point) estimation:** Construct a *function* of \((Y_1, \ldots, Y_n)\) (i.e., a *statistic*) that, if evaluated at a *particular realization* \((y_1, \ldots, y_n)\), yields a numerical value that gives information on the *true value* of \(\psi\)

- **Estimator:** The function itself
- **Estimate:** The numerical value based on a particular realization
- **Estimation:** Used both to denote the procedure (estimator) and actual calculation of a numerical value (estimate)
- Also depend on \(U\) and \(u\), of course

**Remark:** The distinction between estimator and estimate is often subject to *abuse* in the literature

- Meaning is often clear from the context
Simplest example: Suppose that \( f(t, U, \theta) = \theta_1 + \theta_2 t \), \( \theta = (\theta_1, \theta_2)^T \)

- *Simple linear regression model*
- No \( U \) here so *suppress conditioning*
- *Usual assumption* — \( Y_j \sim \mathcal{N}(\theta_1 + \theta_2 t_j, \sigma^2) \), \( Y_1, \ldots, Y_n \) all \( \perp \perp \)

\[ \psi = (\theta^T, \sigma^2)^T \]

**Standard estimator for \( \theta \):** The *ordinary least squares* estimator for \( \theta \)

\[
\arg\min_{\theta} \sum_{j=1}^{n} (Y_j - \theta_1 - \theta_2 t_j)^2
\]

(4)

- For a *particular data set*, the *estimate* of \( \theta \) minimizes

\[
\sum_{j=1}^{n} (y_j - \theta_1 - \theta_2 t_j)^2
\]

- The rest of \( \psi, \sigma^2 \), can be estimated separately (later)
- Note that (4) is the usual objective function adopted in *inverse problems* for arbitrary \( f(t, U, \theta) \)
Closed form solution: Defining

\[ X = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \]

and minimizing yields the estimator and estimate

\[ \hat{\theta}(Y) = (X^T X)^T X^T Y \quad \text{and} \quad \hat{\theta} = (X^T X)^T X^T y \]

- **Convention**: “\(^\wedge\)” means estimator or estimate (for \(\theta\) here)
- **Emphasis** that the estimator \(\hat{\theta}(Y)\) is a function of \((Y_1, \ldots, Y_n)\) is usually suppressed
Motivation for ordinary least squares: Coming up

- Separate estimation of $\sigma^2$: 
  $$\hat{\sigma}^2 = (n - 2)^{-1} \sum_{j=1}^{n} (Y_j - \hat{\theta}_1 - \hat{\theta}_2 t_j)^2$$

- $\hat{\psi} = (\hat{\theta}^T, \hat{\sigma}^2)^T$

**Question:** How “good” is $\hat{\theta} = \hat{\theta}(Y)$ as an estimator for the true value of $\theta$?

- More generally, how good is $\hat{\psi} = \hat{\psi}(Y)$ as an estimator for the true value of $\psi$?

- A question about the *procedure, not* a particular numerical value

- What do we mean by “good?” Can we characterize the *extent of uncertainty*?
**Statistical inference**

**Key idea:**

- An *estimator* is a *function* of random variables that represent the *data generating mechanism*.

- Thus, for any value of $\psi$, the *estimator itself* has *probability distribution* (that depends on that of $(Y_1, \ldots, Y_n)$ and hence on $\psi$).

- In general problems, this distribution is *conditional on $U$*.

- It is typical to suppress this dependence for brevity in general arguments; it is understood that everything is *conditional on $U$* (we’ll do that here).
Statistical inference

Conceptually:

- **Each possible realization** of \((Y_1, \ldots, Y_n)\) yields a (numerical) value of \(\hat{\psi}\); i.e., the random vector \(\hat{\psi}(Y)\) has sample space comprising all these values.

- We may thus think of the **probability distribution** of \(\hat{\psi}(Y)\) as representing probabilities for values of \(\hat{\psi}\) that would be observed across “all possible data sets” we could “end up with”.

- When we **collect data**, we end up with **only one** of these.

- If this distribution has **large variance**, estimates **vary a lot** across possible data sets, and our one estimate tells us little about the true value of \(\psi\) (another data set might have yielded a very different numerical value) ⇒ lots of uncertainty.

- On the other hand, if this distribution has **small variance**, estimates **vary little** across possible data sets, and our one estimate may be quite informative about the true value of \(\psi\) (another data set might have yielded a similar value) ⇒ only mild uncertainty.
Statistical inference

**Sampling distribution:** The *probability distribution* of an *estimator*

- *Properties* of the sampling distribution characterize the *uncertainty* in the *estimation procedure* (*estimator*)

**Unbiasedness:** Intuitively, for a particular $\psi$ (or its components)

- Some values in the sample space are larger than $\psi$, some smaller
- All possible values of *estimator* for $\psi$ should "average out" to $\psi$
- Formally, writing $E_\psi(\cdot)$ to denote expectation with respect to the distribution of $(Y_1, \ldots, Y_n)$ under $\psi$

$$E_\psi(\hat{\psi}) = \psi \quad (5)$$

- An *estimator* that satisfies (5) is called an *unbiased estimator*
Statistical inference

**Sampling variance and standard error:** Quantify *variation* across possible values of $\psi$

- The *sampling covariance matrix* is the covariance matrix of the *sampling distribution* of $\hat{\psi}(Y)$, which we write as
  \[
  \text{var}_\psi(\hat{\psi})
  \]

- Diagonal elements of $\text{var}_\psi(\hat{\psi})$ are the *sampling variances* of the components of $\hat{\psi}(Y)$, e.g., $\text{var}(\hat{\psi}_k)$ for the $k$th component

- The *standard error* is the *standard deviation* of the $k$th component of $\hat{\psi}(Y)$, $\sqrt{\text{var}(\hat{\psi}_k)}$ (on the same scale as $\psi$)

**Key factors determining magnitude of sampling variance:**

- Variance of *original probability distribution* of $Y$ (very well *out of our control*)

- *Sample size* $n$ (often *under our control – experimental design*)
Example, continued: *Sampling distribution of ordinary least squares estimator* (a component of $\psi$)

- The estimator is *unbiased*: $E(Y) = X\theta$

  $$E(\hat{\theta}) = E\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T X\theta = \theta$$

- The *sampling variance*: $\text{var}(Y) = \sigma^2 I_n$

  $$\text{var}(\hat{\theta}) = \text{var}\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T \text{var}(Y) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

- The *sampling distribution* itself is *multivariate normal*:

  $$\hat{\theta} \sim \mathcal{N}_2\{\theta, \sigma^2 (X^T X)^{-1}\}$$

  - *Sampling variance* depends on $\sigma^2$ (variance of the assumed probability distribution of $Y_j$)
  - Writing $(X^T X)^{-1} = n^{-1}\{n^{-1} X^T X\}^{-1}$, *sampling variance decreases* as $n$ *increases*
**Statistical inference**

**Absolute necessity:** *Whenever* an estimate based on data is reported, it should be *accompanied by* an assessment of *uncertainty* based on the *sampling distribution*

- **Natural approach:** look at the *standard error*; here, for the $k$th component of $\hat{\theta}$, $\sqrt{\text{var}(\hat{\theta}_k)} = \sqrt{\sigma^2 (X^T X)^{-1}_{kk}}$

- $\sigma^2$ is not *known*, so provide an *estimate* of the *standard error* of the $k$th component (usually itself called “*standard error*”)

- In the example, $SE(\hat{\theta}_k) = \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{kk}}$, $k = 1, 2$

- **How to interpret?**
  
  (i) If $\hat{\theta}_2 = 2.0$ and $SE(\hat{\theta}_2) = 3.0 \Rightarrow$ very uncertain
  (ii) If $\hat{\theta}_2 = 2.0$ and $SE(\hat{\theta}_2) = 0.03 \Rightarrow$ feeling pretty good!

- **Ways to improve precision (reduce uncertainty)** if case (i)?
**Statistical inference**

**In fact:** We can do more: The *entire sampling distribution* can help us to make a *probability statement* to better characterize *uncertainty*

- For example,

\[
Z = \frac{\hat{\theta}_k - \theta_k}{\sqrt{\sigma^2(X^T X)^{-1}}_{kk}} \sim \mathcal{N}(0, 1) \quad \text{"standard normal distribution,"} \quad k = 1, 2
\]

- When \( \sigma^2 \) is replaced by \( \hat{\sigma}^2 \),

\[
T = \frac{\hat{\theta}_k - \theta_k}{SE(\hat{\theta}_k)} \sim t_{n-2}
\]

*"Student’s t distribution with \( n - 2 \) degrees of freedom"

- Let \( \alpha = \text{some small probability} \), so \( 1 - \alpha \) is large; e.g., \( \alpha = 0.05 \)

- Let \( t_{1-\alpha/2} \) satisfy \( P(T \geq t_{1-\alpha/2}) = \alpha/2 \), so, by *symmetry*,

\[
P(T \leq -t_{1-\alpha/2}) = \alpha/2 \quad \Rightarrow \quad P(-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2}) = 1 - \alpha
\]

\[
P\{\hat{\theta}_2 - t_{1-\alpha/2}SE(\hat{\theta}_2) \leq \theta_2 \leq \hat{\theta}_2 + t_{1-\alpha/2}SE(\hat{\theta}_2)\} = 1 - \alpha
\]
Statistical inference

\[ P\{\hat{\theta}_2 - t_{1-\alpha/2}SE(\hat{\theta}_2) \leq \theta_2 \leq \hat{\theta}_2 + t_{1-\alpha/2}SE(\hat{\theta}_2)\} = 1 - \alpha \]

Important:

- A probability pertaining to the *sampling distribution* of \( \hat{\theta}_2 \)
- That is, for “*all possible realizations of \( Y \) of size \( n \),” probability is \( 1 - \alpha \) that the *endpoints* of the *interval*

\[ [\hat{\theta}_2 - t_{1-\alpha/2}SE(\hat{\theta}_2), \hat{\theta}_2 + t_{1-\alpha/2}SE(\hat{\theta}_2)] \]

will include (the *fixed value*) \( \theta_2 \)

- *Confidence interval*
**Statistical inference**

**Confidence interval:** A probability statement about the *procedure* by which an *estimator* is constructed from a realization of \( Y \)

- *NOT* a probability statement about \( \theta_2 \) (*fixed*)

- **Interpretation:** “For all possible realizations of \( Y \) of size \( n \), if we were to calculate the interval according to the *estimation procedure*, \((1 - \alpha)\%\) of such intervals would ‘cover’ \( \theta_2 \)”

- Provides *more information* than just *standard error* alone – how “large” or “small” SE must be *relative to* \( \hat{\theta}_2 \) to feel “confident” that the procedure of data generation and estimation provides a reliable understanding (“confident” quantified by \( \alpha \))

- \( \alpha \) is chosen by the *analyst*

- For the example, \( \alpha = 0.05 \), \( n \) large \( \Rightarrow \) \( t_{1-\alpha/2} \approx 1.96 \)
  
  (i) \( \hat{\theta}_2 = 2.0, \ SE(\hat{\theta}_2) = 3.0 \) gives \([-3.88,7.88] \) \( \Rightarrow \) *no confidence*

  (ii) \( \hat{\theta}_2 = 2.0, \ SE(\hat{\theta}_2) = 0.03 \) gives \([1.94,2.09] \) \( \Rightarrow \) *feeling pretty confident!*
Warning: The numerical values themselves are meaningless except for the impression they give about the quality of the procedure

- Once a realization of data is in hand, the numerical interval either “covers” the true value of $\theta_2$ or it doesn’t
- Wrong interpretation: The probability is $1 - \alpha$ that $\theta_2$ is between $-3.88$ and $7.88$.
- What we CAN say: We are $(1 - \alpha)\%$ “confident” that intervals constructed this way would “cover” $\theta_2 \Rightarrow$ the numerical values give only a sense of the faith we should attach to results of gathering data the way we did
Recall: In our *modeling context*

- A *statistical model* describes *all possible probability distributions* for $Y$ (conditional on $U$), *indexed by a parameter* $\psi$ (under some set of *assumptions*)

**Statistical inference, formalized:** *Estimation* of $\psi$ and associated assessment of *uncertainty* (*sampling distribution*)

**Issues:**

(i) Derivation of “*sensible*” *estimators* for $\psi$?

(ii) *Competing* estimators? “*Optimality*?”

(iii) *Sensitivity* to possibly *incorrect assumptions*?

(iv) Derivation of *sampling distribution* in *complex* models? (next section)
Example, continued: In the *theophylline example* we assumed

\[ Y_j = f(t_j, U, \theta) + (\epsilon_{1j} + \epsilon_{2j}), \quad j = 1, \ldots, n, \quad \Rightarrow \quad Y = f(U, \theta) + (\epsilon_1 + \epsilon_2) \]

\[ E(\epsilon_1 | U) = 0, \quad E(\epsilon_2 | U) = 0, \quad \text{var}(\epsilon_1 | U) = \sigma_1^2 I_n, \quad \text{var}(\epsilon_2 | U) = \sigma_2^2 I_n \]

and \( \epsilon_1 \perp \perp \epsilon_2 \), so that

\[ E(\epsilon | U) = 0, \quad \text{var}(\epsilon | U) = \sigma^2 I_n, \quad \sigma^2 = \sigma_1^2 + \sigma_2^2 \]

and thus

\[ E(Y | U) = f(U, \theta), \quad \text{var}(Y | U) = \sigma^2 I_n \quad (6) \]

- (6) is an *assumption* on *certain features* of the probability distribution of \( Y|U \) [first two moments, parameterized by \( \psi = (\theta^T, \sigma^2)^T \)]

- *In addition*, we assumed that \( \epsilon|U \) is a *multivariate normally distributed random vector*, and hence so is \( Y|U \), with *mean* and *covariance matrix* given in (6).
**Parametric vs. semiparametric models**

**Parametric statistical model:** A statistical model in which the *probability distribution* of $Y$ is *completely specified*

- For theophylline

\[ Y|U \sim \mathcal{N}_n\{f(U, \theta), \sigma^2 I_n\} \]

- \[ \Rightarrow \] *All* possible probability distributions are *multivariate normal* with this mean and covariance matrix (over all $\psi$)

- *Under this assumption*, all aspects of $Y$ (given $U$) (*probabilities* of any event, all *higher moments*) are (in principle) *known* if we know $\psi$

- For (fully) *parametric models*, there is a *standard approach* to inference (coming up)

- But taking $Y$ is *multivariate normal* is an *assumption* – how will inference be affected if it’s *wrong*?
**Parametric vs. semiparametric models**

**Semiparametric statistical model:** A statistical model in which the probability distribution of $Y$ is only partially specified

- For theophylline

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

*First two moments* in terms of *parameter* $\psi$

- We could *stop here* – this is *all* we’re willing to say

- $\Rightarrow$ *All* possible probability distributions are *any* probability distributions having this mean and covariance matrix (over all $\psi$)

- *Intuition* – fewer *assumptions* $\Leftrightarrow$ including *more* possible models, less chance for *problems* due to *misspecification*

- *Price to pay* – Fewer *assumptions* $\Leftrightarrow$ cannot “*take advantage*” of *additional information* in full probability model if it’s *correct*
Parametric vs. semiparametric models

Estimators in parametric models: Several possibilities

The likelihood function: Suppose \( p(y \mid \psi) \) is the pmf/pdf for \( Y \) under the assumed parametric model (suppress \( U \))

- The notation emphasizes that \( p(y) \) is indexed by \( \psi \) (and is a function of \( y \) for fixed \( \psi \))
- Given that \( Y = y \) is observed, the function of \( \psi \) defined by

\[
L(\psi \mid y) = p(y \mid \psi)
\]

is called the likelihood function

- This is a function of \( \psi \) for fixed \( y \)

Remark: The likelihood function forms the basis for one of the most popular and defensible general strategies for finding estimators

- The maximum likelihood estimator
**Parametric vs. semiparametric models**

**Demonstration:** $Y$ discrete random vector

- $L(\psi \mid y) = p(y \mid \psi) = P_\psi(Y = y)$, where the subscript emphasizes that the pmf is indexed by $\psi$

- Suppose $\psi_1$ and $\psi_2$ are two possible values for $\psi$ such that
  
  $P_{\psi_1}(Y = y) = L(\psi_1 \mid y) > L(\psi_2 \mid y) = P_{\psi_2}(Y = y)$

- **Interpretation** – The $y$ we actually observed is more likely to have occurred if $\psi = \psi_1$ than if $\psi = \psi_2$ ⇒ $\psi_1$ is a “more plausible” value for the true value of $\psi$ than is $\psi_2$

- ⇒ Seems reasonable to examine the probability of the $y$ we actually observed under different possible $\psi$

- Reasoning extends to continuous random vectors under some technicalities
Likelihood principle: If $y_1$ and $y_2$ are two possible realizations of $Y$ such that

$$L(\psi \mid y_1) = C(y_1, y_2)L(\psi \mid y_2)$$

for all $\psi$, for constant $C(y_1, y_2)$ not depending on $\psi$, then conclusions drawn from observing $y_1$ or $y_2$ should be identical.

- May view the likelihood function as a data reduction device – the likelihood function contains all information about $\psi$ in a sample.
- $C(y_1, y_2) = 1 \Rightarrow y_1$ and $y_2$ contain the same information about $\psi$.
- Moreover, if the likelihood functions are only proportional, the information is still the same $\Rightarrow$

$$L(\psi_2 \mid y_1) = 3L(\psi_1 \mid y_1) \Leftrightarrow L(\psi_2 \mid y_2) = 3L(\psi_1 \mid y_2)$$

and will conclude $\psi_2$ is more plausible than $\psi_1$ whether we observe $y_1$ or $y_2$. 


**Parametric vs. semiparametric models**

**Definition:** For each \( y \in Y \), let \( \hat{\psi}(y) \) be a parameter value where \( L(\psi \mid y) \) attains its maximum as a *function* of \( \psi \), with \( y \) held fixed. A *maximum likelihood estimator* for \( \psi \) is \( \hat{\psi}(Y) \)

- Acronym – MLE
- *Intuitively reasonable* – MLE is the *parameter value* for which the observed \( y \) is most likely
- *More formally* – MLE has *optimality properties* under the *assumption* that the specified probability model is *correct*
- *Operationally* – must find the *global maximum* (possible *computational issues*)
- If \( L(\psi \mid y) \) is *differentiable* in each component of \( \psi \), possible candidate for MLE is the value of \( \psi \) satisfying

  \[
  \frac{\partial}{\partial \psi} \log L(\psi \mid y) = 0
  \]

  (Necessary but not sufficient, local maximum, etc)
Parametric vs. semiparametric models

Nice properties:

- **Invariance** – if $\hat{\psi}$ is the MLE for $\psi$, then for any function $\tau(\psi)$, $\tau(\hat{\psi})$ is the MLE for $\tau(\psi)$

- Is the “best” estimator for $\psi$ in a certain sense (coming up)

**Example, continued:** *Simple linear regression model* $Y = (Y_1, \ldots, Y_n)^T$

$Y_j \sim \mathcal{N}(\theta_1 + \theta_2 t_j, \sigma^2)$ are independent, $\psi = (\theta_1, \theta_2, \sigma^2)^T$

- **Under normality assumption**, the likelihood function is

$$L(\psi \mid y) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(y_j - \theta_1 - \theta_2 t_j)^2}{2\sigma^2} \right\}$$

$$\log L(\psi \mid y) = -n \log \sigma - \sum_{j=1}^{n} \frac{(y_j - \theta_1 - \theta_2 t_j)^2}{2\sigma^2}$$
Example, continued:

- Clearly maximized in $\theta_1, \theta_2$ by minimizing
  \[
  \sum_{j=1}^{n}(y_j - \theta_1 - \theta_2 t_j)^2 \implies \text{estimator minimizes} \sum_{j=1}^{n}(Y_j - \theta_1 - \theta_2 t_j)^2
  \]

- Differentiating $\log L(\psi | y)$ wrt $\theta_1, \theta_2$ yields
  \[
  \sum_{j=1}^{n}(y_j - \theta_1 - \theta_2 t_j) \begin{pmatrix} 1 \\ t_j \end{pmatrix} = 0 \tag{7}
  \]
  
  The so-called normal equations

- Differentiating $\log L(\psi | y)$ wrt $\sigma^2$ yields
  \[
  \sigma^2 = n^{-1} \sum_{j=1}^{n}(y_j - \theta_1 - \theta_2 t_j)^2
  \]
Parametric vs. semiparametric models

Result: The ordinary least squares estimator is the maximum likelihood estimator under the assumption of normality.

Inference in semiparametric models: Maximum likelihood seems infeasible without a full distributional assumption.

- Derive estimators in other (sometimes seemingly ad hoc) ways.
**Example, continued: Simple linear regression model**

- **Only willing to assume** $E(Y_j) = \theta_1 + \theta_2 t_j, \quad \text{var}(Y_j) = \sigma^2$
- **Perspective** – “minimize distance between model and data”

$$\text{arg min}_\theta \sum_{j=1}^{n} (Y_j - \theta_1 - \theta_2 t_j)^2 \quad \Rightarrow \quad \text{ordinary least squares estimator}$$

- Alternatively, (7) can be written as

$$n^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ t_j \end{pmatrix} = n^{-1} \sum_{j=1}^{n} (\theta_1 + \theta_2 t_j) \begin{pmatrix} 1 \\ t_j \end{pmatrix}$$

*Equate the LHS to its expectation – “moment method”*

- Solving (7) turns out to yield the “best” estimator for $\theta$ in a certain sense if *all* one is willing specify is the mean and variance of $Y_j$
Remarks:

- **Identifiability of statistical models.** A parameter $\psi$ for a family of probability distributions is *identifiable* if *distinct* values of $\psi$ correspond to *distinct* pmf/pdfs. I.e., if the pmf/pdf for $Y$ is $p(y | \psi)$ under the model

  $$\psi \neq \psi' \Leftrightarrow p(y | \psi) \text{ is not the same function of } y \text{ as } p(y | \psi')$$

- **Identifiability is a feature of the model, NOT of an estimator or estimation procedure**

- **Difficulty with nonidentifiability in inference.** Observations governed by $p(y | \psi)$ and $p(y | \psi')$ look *the same* $\Rightarrow$ no way to know if $\psi$ or $\psi'$ is the true value (same likelihood function value)

- **How to fix nonidentifiability.** May revise model, impose constraints, make assumptions

- **Limitations of data.** Even if a model is identifiable, without sufficient data, it may be *practically impossible* to estimate some parameters, because the information needed is *not available* in the data
Complication: Unfortunately, things aren’t always so easy

- For simple statistical models like the simple linear regression model with normality assumptions, deriving the sampling distribution and showing unbiasedness of the least squares estimator for $\theta$ is straightforward and available exactly

- However, things get ugly quickly; e.g., recall the theophylline example with the parametric model

$$Y_j | U \overset{\text{i.i.d.}}{\sim} \mathcal{N}\{f(t_j, U, \theta), \sigma^2\}$$

$$f(t, U, \theta) = \frac{k_a FD}{V(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}, \; \theta = (k_a, k_e, V)^T$$

- The ordinary least squares estimator minimizes

$$\sum_{j=1}^n \{Y_j - f(t_j, \theta)\}^2 \Rightarrow \text{solve} \; \sum_{j=1}^n \{Y_j - f(t_j, U, \theta)\} f_\theta(t_j, U\theta) = 0$$

$$f_\theta(t, U, \theta) = \frac{\partial}{\partial \theta} f(t, U, \theta)$$
Large sample approximation

• Obviously *not possible* to solve for \( \theta \) *explicitly* \( \Rightarrow \hat{\theta}(Y) \) is a complicated function of \( Y \)

**Result:** Simply moving from a *linear* model to a *nonlinear* (albeit simple) one complicates derivation of the sampling distribution for \( \hat{\theta} \)

• This is the case for *most statistical models*

**Further consideration:** For *semiparametric models*, with no *distributional assumption* for \( Y|U \), even if an *explicit expression* \( \hat{\theta} \) were available, how to derive the full sampling distribution?

**Approach:** Resort to an *approximation*

• *Standard approximation* – evaluate properties of estimators for “large” sample size; i.e., for \( n \to \infty \)

• *Hope* that properties so derived will provide a reasonable approximation to (intractable) properties for realistic (finite) \( n \)
Large sample approximation

Large sample (asymptotic) theory: Evaluate properties as "sample size" → ∞

- Require some basic concepts and tools...

Generically: Let $X_n$, $n = 1, 2, \ldots$ represent a sequence of random variables or vectors indexed by $n$, and let $X$ be a random variable or vector [all on the same probability space $(\Omega, \mathcal{B}, P)$]

Convergence in probability: $X_n$ converges in probability to $X$, written $X_n \xrightarrow{p} X$, if

$$\lim_{n \to \infty} P(\|X_n - X\| < \delta) = 1 \quad \text{for all } \delta > 0$$

- For random vectors, holds iff component-wise convergence

- Fact: If $h(\cdot)$ is continuous, then $X_n \xrightarrow{p} X \Rightarrow h(X_n) \xrightarrow{p} h(X)$
Large sample approximation

**Usefulness:** Identify $X_n$ as a sequence of *estimators* $\hat{\psi}_n$ based on samples of size $n$ for *parameter* $\psi$ in a parametric or semiparametric model

- *Assuming the model is correct*, let $\psi_0$ be the *true value* of $\psi$; i.e, the value that *actually generates* data we might see

- Identify $X$ with $\psi_0$ (a *degenerate* random variable)

**Consistency:** An *estimator* $\hat{\psi}_n$ is said to be *consistent* for $\psi_0$ if

\[ \hat{\psi}_n \xrightarrow{p} \psi_0 \]

**Practical interpretation:** For $n$ “large”, the probability is “small” that $\hat{\psi}_n$ takes value outside an arbitrarily small neighborhood of $\psi_0$

- As $n$ gets “large,” the probability that $\hat{\psi}_n$ will “wander away” from $\psi_0$ is “small”

- Consistency is *good* – sort of like *unbiasedness*
Large sample approximation

Convergence in distribution: Suppose $X_n$ has cdf $F_n$ and $X$ has cdf $F$. $X_n$ is said to converge in distribution (or converge in law) to $X$ iff, for each continuity point of $F$

$$\lim_{n \to \infty} F_n(x) = F(x)$$

Write $X_n \xrightarrow{D} X$ or $X_n \xrightarrow{L} X$

- **Fact** – if $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{L} X$

Practical interpretation: If we are interested in probabilities associated with $X_n$, we can replace them by probabilities for $X$

- Use to approximate sampling distribution of an estimator

Not so fast...

- $\hat{\psi}_n \xrightarrow{p} \psi_0$ (*consistent estimator*) $\Rightarrow$ $\hat{\psi}_n \xrightarrow{L} \psi_0$

- $\psi_0$ is a constant, so the “limit distribution” is degenerate $\Rightarrow$ not particularly useful!
Large sample approximation

**Instead:** Consider suitably *centered and scaled* version of $\hat{\psi}_n$ with “more interesting” behavior

**Asymptotic normality (random variable):** $X_n$ is *asymptotically normal* if we can find sequences $\{a_n\}$ and $\{c_n\}$ such that

$$c_n(X_n - a_n) \xrightarrow{L} \mathcal{N}(0, 1)$$  \hspace{1cm} (8)

- **Notational convention:** this means “converges in law to a random variable with a standard normal distribution”

- **Advantage:** (8) implies $X_n \sim \mathcal{N}(a_n, c_n^2)$, where “$\sim$” means “distributed approximately as”

**For “regular” estimators:** It may be shown that

$$n^{1/2}(\hat{\psi}_n - \psi_0) \xrightarrow{L} \mathcal{N}(0, C),$$

where $C = \lim_{n \to \infty} C_n$ so that

$$\hat{\psi}_n \sim \mathcal{N}(\psi_0, n^{-1}C_n)$$
Large sample approximation

Asymptotic relative efficiency (ARE): Suppose we have two competing estimators $\hat{\psi}_n^{(1)}$ and $\hat{\psi}_n^{(2)}$ satisfying

$$n^{1/2}(\hat{\psi}_n^{(1)} - \psi_0) \xrightarrow{L} \mathcal{N}(0, C_1), \quad n^{1/2}(\hat{\psi}_n^{(2)} - \psi_0) \xrightarrow{L} \mathcal{N}(0, C_2)$$

- If $\text{dim } \psi = 1$, asymptotic relative efficiency of $\hat{\psi}_n^{(1)}$ to $\hat{\psi}_n^{(2)}$

$$\text{ARE} = \frac{C_2}{C_1},$$

and $\hat{\psi}^{(1)}$ is preferred if $\text{ARE} > 1$

- If $\text{dim } \psi = p > 1$, then $\text{ARE} = (|C_2|/|C_1|)^{1/2}$

- If $\text{ARE} > 1$ $\hat{\psi}^{(2)}$, needs proportionately $\text{ARE}$ as many observations as $\hat{\psi}^{(1)}$ to achieve the same (large-$n$) precision
Large sample approximation

Maximum likelihood estimator: For a parametric model, if the model assumptions are all exactly correct, then the maximum likelihood estimator for $\psi$ is asymptotically efficient

- The large-$n$ distribution for a MLE achieves the “smallest” possible variance among all “regular” estimators