The regression likelihood for individual $i$ was:

$$
\text{likelihood}_i = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) e^{-\frac{1}{2\sigma^2}(Y_i - X_i\alpha)^\prime(Y_i - X_i\alpha)}.
$$

For correlated data, when $Y_i$ is a vector with $J$ elements, we will do exactly the same thing: write down a likelihood and maximize it. However, the formula for the bell curve looks a bit different:

$$
\text{likelihood}_i = \left( \frac{1}{\sqrt{2\pi}} \right)^J |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(Y_i - X_i\alpha)^\prime \Sigma_i^{-1}(Y_i - X_i\alpha)}.
$$

Because we assume clusters (e.g., individuals in a longitudinal study, nursing homes in a clustered study) are independent of each other, we multiply together all of the likelihood, across all clusters, take the log, and then maximize, just like in regression.

What exactly is that $|\Sigma_i|^{-\frac{1}{2}}$?? We can also write this as

$$
\frac{1}{\sqrt{|\Sigma_i|}}.
$$

The vertical bars indicate that we compute the determinant of $\Sigma_i$, which is just a number; then we take the square root, and then the inverse. This is the multivariate (multi-dimensional) equivalent of $\frac{1}{\sqrt{\sigma^2}}$ from the 1-dimensional regression likelihood. The determinant is a complex summary measure of the variability in the data.
Back to estimation... consider notation of $\Sigma \equiv \Sigma(\theta)$ to make explicit that there are variance/covariance parameters inside of $\Sigma$ (denoted by $\theta$) which must be estimated.

If we maximize our likelihood for $\alpha$, we get:

$$
\hat{\alpha}(\Sigma) = \left( \mathbf{X}' \Sigma^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Y}.
$$

What does this look like? (Think about weighted regression...) However, we don’t actually know $\Sigma$ so we must estimate it as well and plug it into our formula:

$$
\hat{\alpha} \equiv \hat{\alpha}(\hat{\Sigma}) = \left( \mathbf{X}' \hat{\Sigma}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \hat{\Sigma}^{-1} \mathbf{Y}.
$$

$\hat{\alpha}$ is called the maximum likelihood estimate (MLE) for $\alpha$. This can be written equivalently as:

$$
\hat{\alpha} = \left( \sum_{i=1}^{n} \mathbf{X}'_i \hat{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}'_i \hat{\Sigma}_i^{-1} \mathbf{Y}_i.
$$

Some theoretical facts about $\hat{\alpha}$.

We saw that

$$
\hat{\alpha}(\Sigma) = \left( \mathbf{X}' \Sigma^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Y}
$$

so $\hat{\alpha}(\Sigma)$ is just a complicated linear combination of the $\mathbf{Y}_i$ values. Thus

$$
\hat{\alpha}(\Sigma) \sim N(\alpha, (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1}) \quad (\text{if } F = N) \quad (1)
$$

$$
E[\hat{\alpha}(\Sigma)] = \alpha \quad \text{(unbiased)} \quad (2)
$$

$$
\text{Var}[\hat{\alpha}(\Sigma)] = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \quad (3)
$$

(2) and (3) are straightforward to show using matrix notation. (Try it.)

What about $\hat{\Sigma}$?

- $\hat{\Sigma}$ cannot usually be written down as a closed-form equation like $\hat{\alpha}$ can.
- $\hat{\Sigma}$ is found numerically, that is, by running an algorithm which finds the maximum likelihood for $\Sigma$ through a sophisticated computational search. There are two standard algorithms, ML and REML, which we will discuss later with mixed models. For the estimating equations approach, we will see a consistent estimator is sufficient, e.g., method of moments.
- This search can be slow, because we need both the determinant and the inverse of a matrix (which is of size at least $J \times J$) to evaluate the log likelihood. SAS uses something called a ridge-stabilized Newton-Raphson algorithm.

Note that $\Sigma$ (not $\hat{\Sigma}$) is in those equations. When $\Sigma$ is known, $\hat{\alpha}(\Sigma)$ has this distribution and is called the Best Linear Unbiased Estimator (BLUE).

Since $\Sigma$ is estimated, not known, we really have

$$
\hat{\alpha} \approx \text{N}(\alpha, (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1})
$$

where the approximation gets better as $n$ ($\#$ clusters) gets larger (not as $J$ gets larger!).
Tests and confidence intervals for $\alpha$ will be based on $\text{Var}[\hat{\alpha}]$, where again we have to substitute $\hat{\Sigma}$ in for $\Sigma$:

$$\text{Var}[\hat{\alpha}] = (X'\hat{\Sigma}^{-1}X)^{-1} = \sum_{i=1}^{n}(X'_i\hat{\Sigma}^{-1}X_i)^{-1}.$$  
(taking square root of above will give standard errors)

By just substituting $\hat{\Sigma}$ in for $\Sigma$ in the equation for $\text{Var}[\hat{\alpha}]$, what source of uncertainty are we ignoring?

There are variance estimates which take this into account ("sandwich" estimates) which in SAS can be implemented by using the option EMPIRICAL in the MODEL statement of PROC MIXED. See Diggle, Heagerty, Liang, and Zeger (2002).

---

**ONE POSSIBLE MODELING STRATEGY - OVERVIEW**

- Create summary statistics and EDA plots to explore the mean structure (effects of covariates, trends across time, etc.).
- Decide upon a reasonable starting (full) model for $X\alpha$. It is best to start with a large model, so that covariance structures can be compared with all reasonable covariates already taken into account (included in $X\alpha$).
- Create summary statistics and EDA plots to explore the covariance structure.
- Decide upon several reasonable covariance structures to consider for $\Sigma_i$.
- Fit one model for each covariance structure, using the same full model for $X\alpha$ in each. Compare the models to determine the best covariance structure.

- Using the chosen covariance structure, reduce the full model for $X\alpha$ to the most parsimonious model appropriate for the scientific questions. It will likely take several stepwise reductions to get a final model (just like in regression).
- Compare your full model for $X\alpha$ to your final model for $X\alpha$.
  - If they are not significantly different, breathe a sigh of relief.
  - If they are, then likely there is collinearity among the covariates. Reconsider your stepwise reductions.
- Carry out model diagnostics.

Then again, any change to the model and the question being answered is different.

- Model building approaches would then tend to determine the question being answered as opposed to establishing the question before hand.
TESTING AND INference FOR $\alpha$

Tests and confidence intervals are formally constructed from $\hat{\alpha}$ and $\text{Var}[\hat{\alpha}]$.

EXAMPLE Contact Time Study

Consider a model with three categorical effects for gender, age, shape, and age*shape, and a linear effect for trial (time).

SAS's $\alpha$ vector is

$$\begin{bmatrix}
\alpha_0 \\
\alpha_{\text{female}} \\
\alpha_{\text{male}} \\
\alpha_{\text{child}} \\
\alpha_{\text{yngadult}} \\
\alpha_{\text{adult}} \\
\alpha_{\text{box}} \\
\alpha_{\text{circle}} \\
\alpha_{\text{child} \times \text{box}} \\
\alpha_{\text{yngadult} \times \text{box}} \\
\alpha_{\text{adult} \times \text{box}} \\
\alpha_{\text{child} \times \text{circle}} \\
\alpha_{\text{yngadult} \times \text{circle}} \\
\alpha_{\text{adult} \times \text{circle}} \\
\alpha_{\text{trial}}
\end{bmatrix}$$

as shown on the next pages.
SAS thinks of each of these contrast hypotheses as a matrix equation: 

$$H_0 : B'\alpha = 0$$ 

where the matrix $B'$ contains the contrast coefficients in the CONTRAST statements.

If the $B'$ matrix is a single row (e.g., gender contrast), then there is a t-test for $H_0 : B'\alpha = 0$:

$$t^* = \frac{B'\hat{\alpha}}{\sqrt{B'\text{Var}[\hat{\alpha}]B}} = \frac{B'\hat{\alpha}}{\sqrt{B'(X'\hat{\Sigma}^{-1}X)^{-1}B}}$$

Because of the approximation of the variance in the denominator...

1. this is rarely (never) an exact test;
2. the error or denominator degrees of freedom $\nu$ must be estimated.

SAS can compute five different ways to estimate the degrees of freedom. (We’ll come back to this later.) In general, use the SAS defaults.

The corresponding (1-$a$)100% Confidence Interval is

$$B'\hat{\alpha} \pm t_{\nu,1-a/2} \sqrt{B'(X'\hat{\Sigma}^{-1}X)^{-1}B}$$

or use $Z_{1-a/2}$ instead of $t_{\nu,1-a/2}$.

When the $B'$ matrix has several rows (e.g., age contrast), then there is an F-test for $H_0 : B'\alpha = 0$:

$$F^* = \frac{(B'\hat{\alpha})' [B'(X'\hat{\Sigma}^{-1}X)^{-1}B]^{-1} (B'\hat{\alpha})}{\text{rank}(B)}$$

- The numerator degrees of freedom are $\text{rank}(B)$.
- The denominator degrees of freedom (which correspond to the $[B'(X'\hat{\Sigma}^{-1}X)^{-1}B]^{-1}$ part of the equation) must be estimated, just like for the t-test.

For either the t-test or the F-test,

- if your data set is large enough, so that all denominator degrees of freedom are about 25 or more, then it won’t matter which estimation technique is used.

When testing one or more terms in the mean model at once, use a likelihood ratio test.

Recall that $\alpha$ is being estimated by maximizing the likelihood:

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n/2} |\Sigma|^{-1/2} e^{-1/2(Y-X\hat{\alpha})'\Sigma^{-1}(Y-X\hat{\alpha})}.$$ 

Once estimated, we can compute the numerical value of the likelihood for that model:

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{(n-J)/2} |\hat{\Sigma}|^{-1/2} e^{-1/2(Y-X\hat{\alpha})'\hat{\Sigma}^{-1}(Y-X\hat{\alpha})}.$$ 

If we change what is in the mean model by dropping some terms (i.e., removing columns from $X$), we can compute a new value for the likelihood.
Provided the covariance structure is not changed, we can carry out a $\chi^2$ test as:

$$T^* = \left( \begin{array}{c} \text{Reduced model} \\ -2 \log \text{likelihood} \end{array} \right) - \left( \begin{array}{c} \text{Full model} \\ -2 \log \text{likelihood} \end{array} \right).$$

The degrees of freedom are computed as

$$r = \left( \# \text{parameters in full model's } \alpha \right) - \left( \# \text{parameters in reduced model's } \alpha \right).$$

IMPORTANT NOTE: The two models must be nested, that is, the reduced model must truly be a reduction of the full model to a smaller model by removing terms. Just like the t-test and the F-test, the likelihood ratio test (LRT) is approximate and the approximation gets better as the sample size $n$ ($\# \text{ clusters}$) gets larger.