Linear Marginal Models - Asymptotic Theory

A marginal linear model is one technique for modelling (continuous) longitudinal data. The idea is to take the concept of a standard (OLS) regression, which makes the normality assumption, and extend this to making a **multivariate normality** assumption. That is, we assume that $Y_i \sim \text{MVN}(X_i\beta, \Sigma_i)$ for some variance structure, Σ_i.

Likelihood Theory

Recall that the density of a multivariate normal is given by

$$f(Y;\mu,\Sigma) = (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)\right).$$

If we assume that $\mu = X_i\beta$, we can derive the MLE for β through standard likelihood arguments.

$$\begin{aligned}
L(\beta,\Sigma_i) &= \prod_{i=1}^{n} f(Y_i;\beta,\Sigma_i) \\
\ell(\beta,\Sigma_i) &= \sum_{i=1}^{n} \log f(Y_i;\beta,\sigma_i) \\
&= \sum_{i=1}^{n} \log\left\{(2\pi)^{-k/2}|\Sigma_i|^{-1/2}\exp\left(-\frac{1}{2}(Y_i-X_i\beta)'\Sigma_i^{-1}(Y_i-X_i\beta)\right)\right\} \\
&= \sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(Y_i-X_i\beta)'\Sigma_i^{-1}(Y_i-X_i\beta) \\
&= -\frac{nk}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{n}(Y_i-X_i\beta)'\Sigma_i^{-1}(Y_i-X_i\beta).
\end{aligned}$$

The last equality follows if we assume that Σ_i is constant for all i.

We can then write down the score function for β as

$$\begin{aligned}
S_\beta &= \frac{\partial}{\partial\beta}\ell(\beta,\Sigma_i) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}(Y_i-X_i\beta)'\Sigma_i^{-1}(Y_i-X_i\beta) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left\{-2X_i'\Sigma_i^{-1}(Y_i-X_i\beta)\right\} \\
&= \sum_{i=1}^{n}X_i\Sigma_i^{-1}Y_i - \sum_{i=1}^{n}X_i\Sigma_i^{-1}X_i\beta.
\end{aligned}$$

The first terms of the likelihood are all free of β, so they differentiate to zero. The actual derivative can be taken using the help of the Matrix Cookbook (78), or through multivariate calculus.

Solving $S_\beta = 0$ for $\widehat{\beta}$ gives us the result that

$$\widehat{\beta} = \left\{\sum_{i=1}^n X_i \Sigma_i^{-1} X_i\right\}^{-1} \sum_{i=1}^n X_i \Sigma_i^{-1} Y_i.$$

A similar process can be followed to derive the MLE for σ^2 (presented as a question on Assignment 1). For ρ, we typically use **profile likelihood**, which occurs by forming the likelihood equation $\ell(\widehat{\beta}, \widehat{\sigma}^2, \rho)$, and maximizing this (numerically) results in an estimate for $\widehat{\rho}$. We can then substitute the estimated $\widehat{\rho}$ back-in to the estimators for $\widehat{\sigma}^2$ and $\widehat{\beta}$ to get numeric values for these.

Finite Sample Bias (Restricted MLE)

If you consider the standard estimator of sample variance, we use (for instance) $\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ instead of the MLE (under normality) $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. The reason is that, while the MLE is asymptotically unbiased and fairly well behaved, it can have appreciable bias in finite samples. The same is true for the MLE's of $\widehat{\rho}$ and $\widehat{\sigma}^2$ presented above. While these estimators are permissible, they are asymptotically unbiased, and will be consistent, if your sample size is not sufficiently large the estimators will be biased.

As a result, it is advised to use **restricted maximum likelihood estimation (REML)** in place of true ML estimation. The idea with REML is that a modified log-likelihood function is optimized, in place of the true log-likelihood, for the purpose of estimating σ^2 and ρ. In particular,

$$\ell_R(\sigma^2, \rho) = \ell(\widehat{\beta}, \sigma^2, \rho) - \frac{1}{2}\log\left|\sum_{i=1}^n X_i' \Sigma_i^{-1} X_i\right|.$$

Otherwise the procedure is essentially equivalent: estimators are selected which maximize this, giving $\widetilde{\sigma}^2$ and $\widetilde{\rho}$, which are then plugged back in to $\widehat{\beta}(\widetilde{\rho})$, giving the REML estimator of β, denoted $\widetilde{\beta}$.

Asymptotic Distribution of $\widehat{\beta}$

Asymptotically, the MLE (and REML) of β follow $N(\beta, \text{var}(\widehat{\beta}))$, where

$$\text{var}(\widehat{\beta}) = \left[\sum_{i=1}^n X_i' \Sigma_i^{-1} X_i\right]^{-1},$$

and which can be estimated by plugging in the corresponding estimates. We can use this asymptotic distribution to perform **Wald type** hypothesis tests, and build Wald type confidence intervals.

In particular, this asymptotic distribution gives us the fact that for every j, $\widehat{\beta}_j \dot{\sim} N(\beta_j, \text{var}(\widehat{\beta})_{j,j})$, where $\text{var}(\widehat{\beta})_{j,j}$ is the j-th diagonal entry of the variance matrix $\text{var}(\widehat{\beta})$. We can then conduct inference based on this distribution. In particular:

$$\begin{aligned} \frac{(\widehat{\beta}_j - \beta_j)^2}{\text{var}(\widehat{\beta})_{j,j}} &\dot{\sim} \chi^2_1 \\ \implies \frac{\widehat{\beta}_j - \beta_j}{\text{s.e.}(\widehat{\beta}_j)} &\dot{\sim} N(0,1). \end{aligned}$$

We can use these results to perform hypothesis tests of $H_0 : \beta_j = \beta^*$, where under the null hypothesis $\frac{\widehat{\beta}_j - \beta^*}{\text{s.e}(\widehat{\beta}_j} \sim N(0,1)$, and so a p-value can be computed as $2P\left(Z > \frac{|\widehat{\beta}_j - \beta^*|}{\text{s.e}(\widehat{\beta}_j)}\right)$. Alternatively, a $100(1-\alpha)\%$ confidence interval can be computed as $\widehat{\beta}_j \pm Z_{\alpha/2}\text{s.e.}(\widehat{\beta}_j)$, where $Z_{\alpha/2}$ is the upper-tail $\alpha/2$ percentile of a $N(0,1)$ random variable.

Asymptotic Distribution of $L\widehat{\beta}$

For the purpose of joint linear hypotheses, or for the purpose of prediction, it is often the case that we wish to test a hypothesis of the form $H_0 : L\beta = \mathbf{c}$ for some matrix L. You can derive (through an argument on quadratic forms) that

$$(L\widehat{\beta} - L\beta)' \left\{L\,\text{var}(\widehat{\beta})L'\right\}^{-1} (L\widehat{\beta} - L\beta) \sim \chi^2_r,$$

where r is the rank of L.

Generally, this result can be used in the same way as the above result, substituting the $N(0,1)$ distribution for the relevant χ^2 distribution. This strategy is used (for instance) to test a set of effects equal to zero simultaneously, or to test a set of effects equal to one another (or equal in ratios to one another). If you take L to be a matrix corresponding to the variate values for an individual you would like to make predictions about, this can also serve as a method of performing inference on predictions.

Parameter Interpretations

The interpretations of parameters in marginal linear models is going to be dictated, in large part, based on the way that time is included in the model. In general, parameter interpretations will take a similar "flavour" to those used in linear regression models. In particular, recall that marginal models are models for the conditional mean of the outcome, given variates ($E[Y_i|X_i]$) and so parameter interpretations will be with regards to the expected change in outcome (based on unit changes in variates, holding all else constant).