Statistical Analysis

An index i refers to family and j to "sibling" within that family. There are n_i siblings in the i^{th} family, $\mathcal{F}_i$; thus, $j = 1, ..., n_i$. The observed phenotype for the j^{th} sibling in the i^{th} family, in an appropriate scale, is y_{ij}.

For OGTT(-10) = fasting glucose, the scale is (fasting glucose) $^{-1.4} \times (-10^5)$. For OGTT (60) and OGTT (120), both less important here, the scale is logarithmic. We focus in what follows on OGTT(-10). In any case, the model is

$$y_{ij} = \mu + f_i + \alpha_{ij} + e_{ij} \ .$$

Here, μ is a constant, the overall mean. The random f_i is the effect of $\mathcal{F}_i$ on each y_{ij} for that i. $E(f_i) \equiv 0$, $\mathrm{Var}(f_i) \equiv \sigma^2_{f_i} \equiv \sigma^2_f$. The "errors" e_{ij} are random; $E(e_{ij}) \equiv 0$, $\mathrm{Var}(e_{ij}) \equiv \sigma^2_e$; α_{ij} is the effect of the genotype of the j^{th} sibling in the i^{th} family on y_{ij}.

Here, the trait is assumed recessive. α_{ij} assumes only two values: $\alpha_1 > 0$ and $\alpha_2 < 0$. Unconditionally, $\{\alpha_{ij}\}$ are taken to be random.

Conditional on the recessive genotype, $\alpha_{ij} = \alpha_1$; conditional on the other genotypes, $\alpha_{ij} = \alpha_2$. We assume that unconditionally, $E(\alpha_{ij}) \equiv 0$. $\mathrm{Var}(e_{ij})$ is then necessarily

$$\sigma^2_\alpha = [P(\alpha_{ij} \text{ recessive}) \times \alpha_1^2] + [(1 - P(\alpha_{ij} \text{ recessive})) \times \alpha_2^2].$$

$\{f_i, e_{ij}, \alpha_{ij} : j = 1, ..., n_i, \ i = 1, 2, ...\}$ are assumed uncorrelated. They are not assumed independent. Certainly $\{f_i, e_{ij}\}$ are not assumed Gaussian, jointly or even marginally.

If $\{f_i, \alpha_i, e_{ij}\}$ were taken to be independent, then it follows from a theorem of H. Cramér that y_{ij} cannot be Gaussian because α_{ij} is not.

Unconditionally, $\mathrm{Var}(y_{ij}) = \sigma^2_y = \mathrm{Var}(f_i) + \mathrm{Var}(\alpha_{ij}) + \mathrm{Var}(e_{ij}) = \sigma^2_f + \sigma^2_\alpha + \sigma^2_e$.

The covariance between y_{ij} and $y_{ij'}$, $\mathrm{Cov}(y_{ij}, y_{ij'}) = \sigma^2_\alpha$; thus, the correlation between the phenotypes of siblings is $\sigma^2_\alpha/(\sigma^2_f + \sigma^2_\alpha + \sigma^2_e)$.

We turn now to estimation, in particular to a "method of moments" approach. Because $E(f_i) \equiv E(\alpha_{ij}) \equiv E(e_{ij}) \equiv 0$, unconditional on genotype $E(y_{ij}) = \mu$. As a consequence, the average of y_{ij} over all siblings and families is unbiased for μ. Call this global average $\hat{\mu}$. Now compute $\{y_{ij} - \hat{\mu}\}$. Of these, n_1, say, will have the recessive genotype and n_2 either

of the other two genotypes. Average the n_1 numbers $\{y_{ij} - \hat{\mu} : \alpha_{ij} = \alpha_1\}$ to obtain the estimate $\hat{\alpha}_1$ of α_1. Average the n_2 numbers $\{y_{ij} - \hat{\mu} : \alpha_{ij} = \alpha_2\}$ to obtain the estimate $\hat{\alpha}_2$ of α_2. Note that $\hat{\mu}, \hat{\alpha}_1$, and $\hat{\alpha}_2$ are computed from sibships of all sizes. Now, fix a family i and siblings j and j'. Compute

$$\begin{aligned}(y_{ij} - y_{ij'})^2 &= ((\alpha_{ij} - \alpha_{ij'}) + (e_{ij} - e_{ij'}))^2 \\ &= (\alpha_{ij} - \alpha_{ij'})^2 + (e_{ij} - e_{ij'}^2) + 2(\alpha_{ij} - \alpha_{ij'})(e_{ij} - e_{ij'}).\end{aligned}$$

Our assumptions on correlations entail that

$$\begin{aligned}\mathrm{Cov}(\alpha_{ij}, e_{ij}) &= \mathrm{Cov}(\alpha_{ij}, e_{ij'}) \\ &= \mathrm{Cov}(\alpha_{ij'}, e_{ij}) \\ &= \mathrm{Cov}(\alpha_{ij'}, e_{ij'}) \\ &= 0\ .\end{aligned}$$

We estimate $(\alpha_{ij} - \alpha_{ij'})^2$ by $(\hat{\alpha}_{ij} - \hat{\alpha}_{ij'})^2$, where $\hat{\alpha}_{ij}$ is $\hat{\alpha}_1$ if α_{ij} is recessive; and $\hat{\alpha}_{ij} = \hat{\alpha}_2$ otherwise; $\hat{\alpha}_{ij'}$ is defined by analogy for sibling j'. Also, $E\{(e_{ij} - e_{ij'})^2\} = E(e_{ij}^2) + E(e_{ij'}^2) - 2E(e_{ij}e_{ij'}) = \sigma_e^2 + \sigma_e^2 - (2 \times 0)$ in view of our assumptions on correlations. It follows from this that

$$(y_{ij} - y_{ij'})^2 - (\hat{\alpha}_{ij} - \hat{\alpha}_{ij'})^2$$

is approximately unbiased for $2\sigma_e^2$. Write N_k for the number of families with k siblings, and assume there are at most K siblings. There are $k(k-1)$ ways of choosing ordered pairs of siblings within $\mathcal{F}_i$, given that it has $k \geq 2$ siblings. It follows that the

$$\max\left\{\frac{1}{2\sum_{k=2}^{K} N_k\ k(k-1)} \sum_i \sum_{(j,j') \in \mathcal{F}_i} \{(y_{ij} - y_{ij'})^2 - (\hat{\alpha}_{ij} - \hat{\alpha}_{ij'})^2\},\ 0\right\}$$

is approximately unbiased for σ_e^2. Call the estimate $\hat{\sigma}_e^2$. This estimate depends primarily on families with at least two siblings, though obviously $\hat{\alpha}_{ij}$ and $\hat{\alpha}_{ij'}$ are computed from all the data. It is inevitable that, with so few assumptions on the components of our model, we will be forced to look mostly within families having at least two siblings to estimate σ_e^2.

Finally, we recall that unconditional on genotype, $\sigma_y^2 = \mathrm{Var}(y_{ij}) = \sigma_f^2 + \sigma_\alpha^2 + \alpha_e^2$. Thus, we estimate σ_f^2 by subtraction. Estimate σ_y^2 by

$$\hat{\sigma}_y^2 = \frac{1}{\sum_{k=2}^{k} N_k} \sum_{i,j} (y_{ij} - \hat{\mu})^2.$$

Estimate σ_α^2 by

$$\hat{\sigma}_\alpha^2 = (\frac{n_1}{n_1+n_2})\hat{\alpha}_1^2 + (\frac{n_2}{n_1+n_2})\hat{\alpha}_2^2 \ ,$$

and σ_f^2 by

$$\hat{\sigma}_f^2 = \hat{\sigma}_y^2 - \hat{\sigma}_\alpha^2 - \hat{\sigma}_e^2 \ .$$

With all this done, σ_y^2 for OGTT(-10) in the given scale is estimated to be made up of 1.76% σ_α^2, 24.81% σ_f^2, and 73.43% σ_e^2. That is, $\hat{\sigma}_\alpha^2/\hat{\sigma}_y^2 = 0.0176$, and so on.