

Two-part estimated means

The default DATA TWOPART approach is to create $\ln Y$ (that is, e-log of Y), for subjects with $Y > 0$ (binary indicator $U=1$) and denote Y as missing for $Y=0$ ($U=0$). The model-estimated means are found in the RESIDUAL output. Note that these are in the $\ln Y$ scale, where Y is the original outcome before DATA TWOPART has been used. Note also that the estimated means are for observations with $Y > 0$. To compute the estimated means in the original scale for all subjects, note that for a normal variable $Z = \ln Y \sim N(m, v)$,

$$E(Y) = \text{Exp}(m+v/2),$$

where m is the mean and v is the variance.

For a two-part model the mean of Y (not taking the e-log and not conditioning on $Y > 0$) is then

$$E(Y) = P(U=0)*0 + P(U=1)*\text{Exp}(m+v/2),$$

where the estimated $P(U=1)$ is found in the section UNIVARIATE DISTRIBUTION FIT.

For example, UG ex 6.16 has model-estimated mean 0.53 for the CONT1 outcome and the estimated $P(U=1)$ is 0.599, resulting in the estimated $E(Y) = 0.599 * \text{Exp}(0.53 + 2.488/2) = 3.53$, where the variance 2.488 for CONT1 is obtained as the sum of the IY variance 1.905 and the CONT1 residual variance 0.583.

The corresponding observed mean can be obtained in a separate TYPE=BASIC run for Y1. In this case it is 4.27.