Formula Guide

Stability Analysis

Stability analysis is the study of how drug product potency degrades over time. The primary statistical quantity of interest is the expiration date, or shelf life. Typically, a drug product is manufactured in batches. When estimating the shelf life of a medication, it is necessary to evaluate how the batches differ with respect to the potency degradation of the drug product over time. Specifically, three different model types are considered.

Model 1: Common Regression across All Batches

This model is appropriate when the intercepts and slopes do not differ across batches and the data therefore can be pooled in order to estimate a common intercept and common slope.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where,

Y_i = response at i^{th} time point
β_0 = common intercept
β_1 = common slope
$x_i = i^{th}$ time point
$\epsilon_i \sim N(0, \sigma^2)$
$i = 1, \ldots, n$
$n =$ total number of observations

Model 2: Separate Intercepts and Common Slope

This model is appropriate when the intercepts differ across batches, but the slopes are common across batches.

$$Y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij}$$

where,

Y_{ij} = response at i^{th} time point and j^{th} batch
β_{0j} = intercept of j^{th} batch
β_1 = common slope across batches
$x_{ij} = i^{th}$ time point for j^{th} batch
$\epsilon_{ij} \sim N(0, \sigma^2)$
$i = 1, \ldots, n_j$
n_j = number of time points for j^{th} batch
$j = 1, \ldots, k$
k = number of batches

Model 3: Separate Intercepts and Separate Slopes.

This model is appropriate when the intercepts and slopes differ across batches. Additionally, the Mean Square Error (MSE) is not pooled across different batches, that is, homogeneity of variance across batches is not assumed.

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

where,

Y_{ij} = response at i^{th} time point and j^{th} batch
β_{0j} = intercept of j^{th} batch
β_{1j} = slope of j^{th} batch
$x_{ij} = i^{th}$ time point for j^{th} batch
$\epsilon_{ij} \sim N(0, \sigma_j^2)$
$i = 1, \ldots, n_j$
n_j = number of time points for j^{th} batch
$j = 1, \ldots, k$
k = number of batches

Parameter Estimation

Parameters are estimated by ordinary least squares. This can be accomplished by solving the normal equations:

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\beta} = \boldsymbol{X}^T Y$$

where,

$\boldsymbol{X}$ = N by p design matrix of coded predictor variables
Y = N by 1 vector of observed values
N = Number of observations
p = Number of columns in design matrix

If the columns of the design matrix are linearly independent then the vector of parameter estimates is as follows:

$$\hat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T Y$$

t-Statistic

The *t*-statistic associated with the null hypothesis that the i[th] parameter is equal to 0 is given by:

$$t = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

where,

$\hat{\sigma}_{\hat{\beta}_i}$ = estimated standard error of $\hat{\beta}_i$
square root of ith diagonal element of MSE $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

Confidence Interval for a Single Parameter

A two-sided 95% confidence interval for the i[th] parameter is given by:

$$\hat{\beta}_i \pm t_{0.975,\mathrm{df}_e}\hat{\sigma}_{\hat{\beta}_i}$$

where,

$t_{0.975,\mathrm{df}_e}$ = 97.5th quantile of a t distribution with df_e degrees of freedom

Predicted Values

The predicted value or estimated mean response for a given covariate vector X_i is given by:

$$\hat{Y}_i = X_i^T \hat{\beta}$$

where,

$$X_i = \left[1\ X_1 \cdots X_{p-1}\right]^T$$

Confidence Interval

The two-sided 95% confidence interval for the mean response for a given covariate vector X_i is given by:

$$\hat{Y}_i \pm t_{0.975,\mathrm{df}_e} \sqrt{MSE\left(X_i^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} X_i\right)}$$

where,

$\mathrm{df}_e =$ error degrees of freedom

$X_i = \left[1\ X_1 \cdots X_{p-1}\right]^T$

$t_{0.975,\mathrm{df}_e} = 97.5^{th}$ quantile of t distribution with $\mathrm{df} = \mathrm{df}_e$

For a one-sided interval, then the 95th quantile of *t* distribution is used.

Prediction Interval

The two-sided 95% prediction interval for a new observation X_i is given by:

$$\hat{Y}_i \pm t_{0.975,\mathrm{df}_e} \sqrt{1 + MSE\left(X_i^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} X_i\right)}$$

where,

$\mathrm{df}_e =$ error degrees of freedom

$X_i = \left[1\ X_1 \cdots X_{p-1}\right]^T$

$t_{0.975,\mathrm{df}_e} = 97.5^{th}$ quantile of t distribution with $\mathrm{df} = \mathrm{df}_e$

For a one-sided interval, then the 95th quantile of t distribution is used.

Model Selection

STATISTICA selects the model by first evaluating if the slopes are equal across different batches.

If the slopes' *p*-value is significant (by default, less than 0.25), the *separate intercepts, separate slopes* model is used to estimate shelf life.

If the slopes' *p*-value is not significant, *STATISTICA* determines if the intercepts are equal across different batches. If the intercepts' *p*-value is significant (by default, less than 0.25), the *separate intercepts, common slope* model is used. If neither the slopes' *p*-value nor the intercepts' *p*-value are significant (by default, less than 0.25) the *common intercept, common slope* model is used.

The *p*-values are computed by using the type 1 sums of squares decomposition of the full linear model with effects placed into the model in the following order: time, batch, batch by time interaction. The slopes' *p*-value is associated with the batch by time interaction and the intercepts' *p*-value is associated with the batch effect.

Shelf Life Estimation

Shelf life estimation is performed by determining where the 95% lower and/or upper confidence or prediction interval intersects with the user-defined specification limit(s). If either M*odel 2* or *Model 3* has been selected, shelf life is estimated separately for each batch. Shelf life is then defined as the shelf life associated with the worst-case batch, that is, the smallest shelf life estimate. By default, *STATISTICA* will base shelf life computations on a two-sided 95% confidence interval.

Residual Analysis

A user can optionally perform a residual analysis to determine the selected model's adequacy.

A **residual** is defined as the difference between the observed and predicted value, that is:

$$e_{ij} = Y_{ij} - \hat{Y}_{ij}$$

where,

Y_{ij} = observed value at i^{th} time point and j^{th} batch
$\hat{Y}_{ij}$ = predicted value at i^{th} time point and j^{th} batch

Additionally, **studentized residuals** are computed. The formula for the studentized residuals is given by:

$$s_{ij} = \frac{e_{ij}}{\sqrt{MSE(1-h_{ii})}}$$

where,

e_{ij} = residual for the i^{th} time point and j^{th} batch
MSE = mean squared error associated with selected model
$h_{ii} = i^{th}$ diagonal element of hat matrix
hat matrix $= \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$
$\boldsymbol{X}$ = design matrix of selected model

A **normal probability plot** of the (studentized) residuals is constructed as follows. First, the (studentized) residuals are rank ordered. From these ranks, *STATISTICA* computes *z* values (i.e., standardized values of the normal distribution) based on the assumption that the data come from a normal distribution

$$z_i = \phi^{-1}\left[\frac{3i-1}{3N+1}\right]$$

where,

ϕ^{-1} is the inverse normal cumulative distribution function
N = Number of observations
$i = 1, \ldots, N$

These *z* values are plotted on the y-axis in the plot. If the observed (studentized) residuals (plotted on the x-axis) are normally distributed, all values should fall onto a straight line. If the residuals are not normally distributed, they will deviate from the line. Outliers can also become evident in this plot. If there is a general lack of fit and the data seem to form a clear pattern (e.g., an S shape) around the line, the variable may have to be transformed in some way (e.g., a log transformation to "pull-in" the tail of the distribution, etc.)