Stability analysis of censored and over-rounded data

Bernhard Schmelzer
Non-clinical statistics conference, Wiesbaden
September 26, 2024

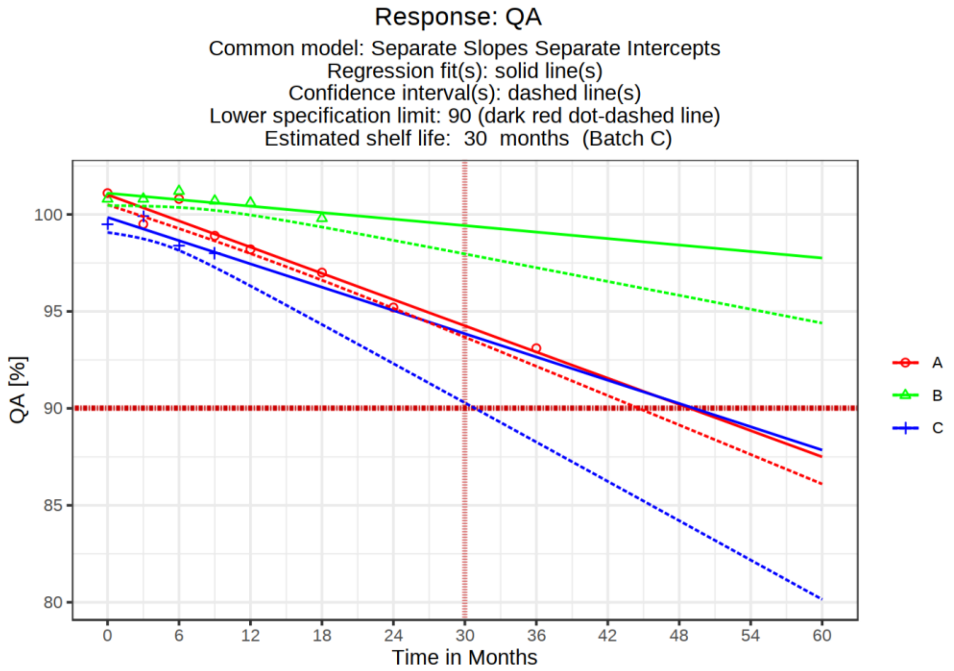
NOVARTIS | Reimagining Medicine

Disclaimer

- These slides are intended for educational purposes only and for the personal use of the audience. These slides are not intended for wider distribution outside the intended purpose without presenter approval.
- The content of this slide deck is accurate to the best of the presenter’s knowledge at the time of production.

Stability analysis and shelf-life computation according to ICH Q1E

- shelf life is the time point at which the 95% confidence limit for the mean response intersect(s) the acceptance/specification limit
- Mean response corresponds to regression line
- 95% confidence limit accounts for the uncertainty in predicting the mean quality / regression line
- Simplest approach: linear change with time
- When data set comprises several batches, it needs to be assessed/tested whether stability behaviour can be described by one common regression line or whether slopes and/or intercepts are significantly different → Analysis of Covariance (ANCOVA) using fixed effects



In this presentation, ANCOVA is out of scope, a fixed effect model assuming common slope for all batches and distinct intercepts will be used.

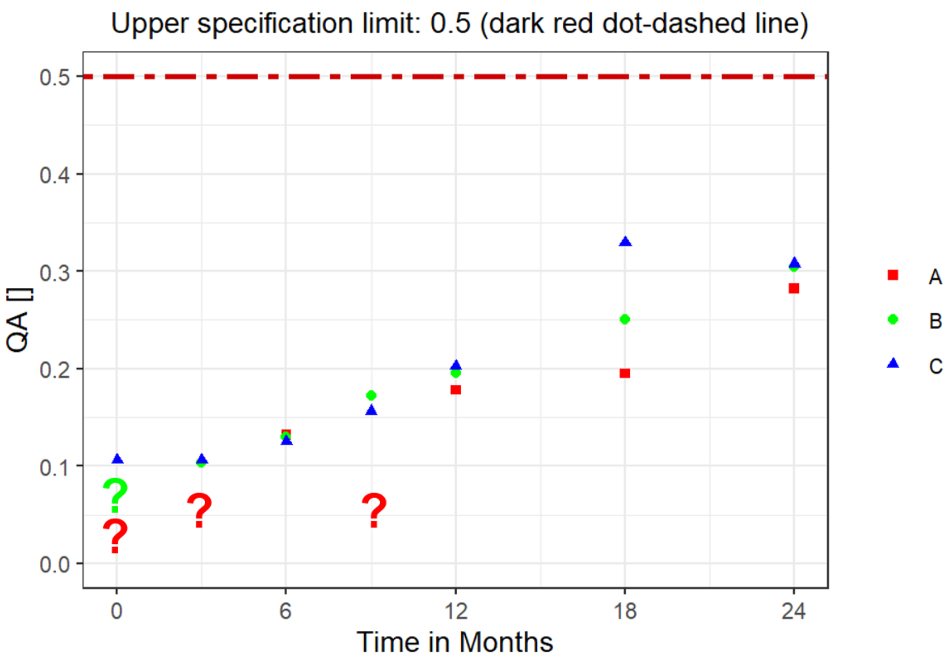
Stability data below limit of quantification (<LOQ)

Frequently, there are quality attributes for which some measurements are below the limit of quantification (<LOQ)

Example: LOQ = 0.095, upper specification limit = 0.5

How to perform stability analysis on such data?

- Omit results <LOQ, treat as missing
- Replacement by fixed value:
  - 0
  - LOQ/2
  - LOQ



batch	time	QA
A	0	<LOQ
A	3	<LOQ
A	6	0.1323
A	9	<LOQ
A	12	0.1780
A	18	0.1951
A	24	0.2820
B	0	<LOQ
B	3	0.1037
B	6	0.1296
B	9	0.1727
B	12	0.1954
B	18	0.2500
B	24	0.3048
C	0	0.1066
C	3	0.1062
C	6	0.1252
C	9	0.1567
C	12	0.2025
C	18	0.3292
C	24	0.3077

Stability data “over-rounded” in addition

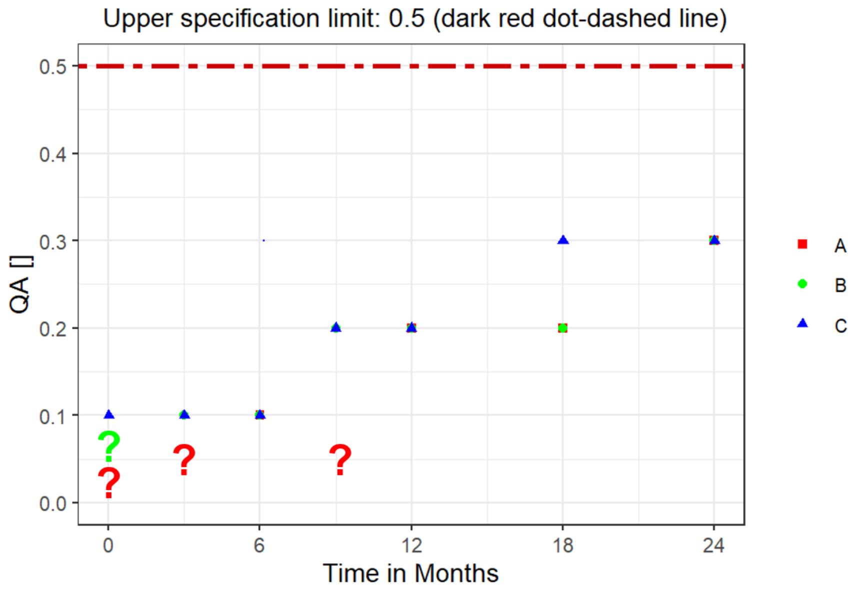
Usually results above LOQ are rounded to the number of digits in the specification limit. This sometimes results in only very few distinct values such that the quality attribute is strongly discretized.

Example: LOQ = 0.095, upper specification limit = 0.5, rounding to one digit after the comma

How to perform stability analysis on such data?

Use rounded results **and** for <LOQ

- Omit results, treat as missing or
- Replacement by fixed value:
  - 0
  - LOQ/2
  - LOQ



batch	time	QA
A	0	<LOQ
A	3	<LOQ
A	6	0.1
A	9	<LOQ
A	12	0.2
A	18	0.2
A	24	0.3
B	0	<LOQ
B	3	0.1
B	6	0.1
B	9	0.2
B	12	0.2
B	18	0.2
B	24	0.3
C	0	0.1
C	3	0.1
C	6	0.1
C	9	0.2
C	12	0.2
C	18	0.3
C	24	0.3

Stability data example – results of replacement approaches

Results using rounded data and <LOQ approaches, compared to using unrounded/uncensored

Stability model: fixed effect model with common/same slope for all batches, different intercepts

Approach	Slope	Worst intercept	Residual SD	Shelf life
Unrounded (raw)	*0.009956*	*0.088173*	*0.026566*	*37.4*
Replace by 0 (repl0)	0.011244	0.070059	0.042981	33.2
Replace by LOQ/2 (replLOQ2)	0.010169	0.081112	0.033738	36.4
Replace by LOQ (replLOQ)	0.009095	0.092164	0.030605	39.5
Treat as missing (missing)	0.009644	0.086516	0.030077	37.4

- Slope over-estimated with repl0, under-estimated with replLOQ, estimate close to the one obtained for unrounded data in case of replLOQ/2
- RMSE over-estimated with all approaches, in particular repl0
- Shelf lives vary quite a lot, shortest shelf life (~4 months shorter) with repl0
- Treat <LOQ values as missing seems to come closest to the shelf life of the unrounded data → general statement hardly possible as this depends on the proportion of data <LOQ

Stability data – interval observations

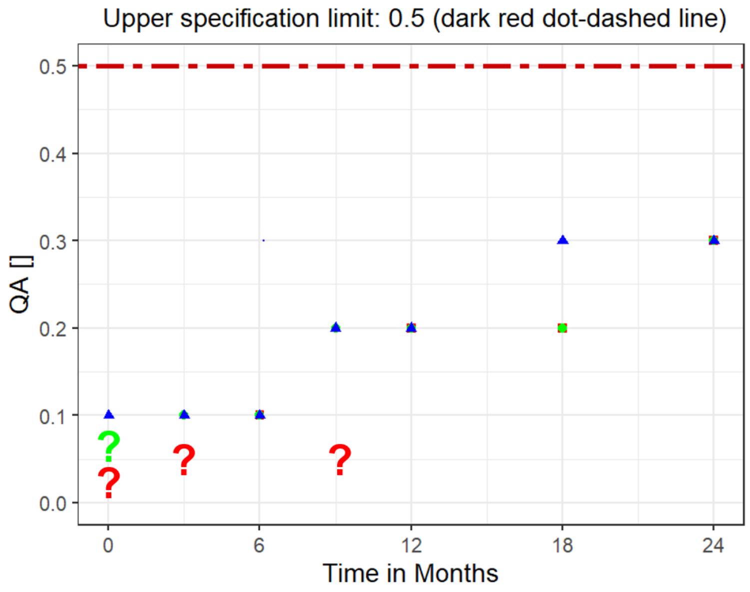
Usually results above LOQ are rounded to the number of digits in the specification limit. This sometimes results in only very few distinct values such that the quality attribute is strongly discretized.

Example: LOQ = 0.095, upper specification limit = 0.5, rounding to one digit after the comma

How to perform stability analysis on such data?

All one can conclude from the reported value is that the true measured value lies within an interval

→ **Treat the data as interval data!**



batch	time	QA	QA_interval
A	0	<LOQ	[0,0.095)
A	3	<LOQ	[0,0.095)
A	6	0.1	[0.095,0.15)
A	9	<LOQ	[0,0.095)
A	12	0.2	[0.15,0.25)
A	18	0.2	[0.15,0.25)
A	24	0.3	[0.25,0.35)
B	0	<LOQ	[0,0.095)
B	3	0.1	[0.095,0.15)
B	6	0.1	[0.095,0.15)
B	9	0.2	[0.15,0.25)
B	12	0.2	[0.15,0.25)
B	18	0.2	[0.15,0.25)
B	24	0.3	[0.25,0.35)
C	0	0.1	[0.095,0.15)
C	3	0.1	[0.095,0.15)
C	6	0.1	[0.095,0.15)
C	9	0.2	[0.15,0.25)
C	12	0.2	[0.15,0.25)
C	18	0.3	[0.25,0.35)
C	24	0.3	[0.25,0.35)

Maximimum likelihood approach

Consider linear fixed effect model $Y = X\beta + \epsilon$

- Y is the response vector (n observations)
- β is the parameter vector containing intercept(s) and slope(s)
- X is the model matrix
- ϵ is the vector of mutually independent, normally distributed random errors $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Given n interval censored observations $\mathbb{Y}_1 = [\underline{y}_1, \overline{y}_1], \ldots, \mathbb{Y}_n = [\underline{y}_n, \overline{y}_n]$
the (observed) likelihood function reads as

$$L(\beta, \sigma^2; \mathbb{Y}_1, \ldots, \mathbb{Y}_n) = \prod_{i=1}^{n} \mathrm{P}(Y_i \in \mathbb{Y}_i) = \prod_{i=1}^{n} \int_{\underline{y}_i}^{\overline{y}_i} g\big(y_i \big| X_{i*}\beta, \sigma^2\big)\, \mathrm{d}y_i$$

with $g\big(\cdot \,\big|\mu, \sigma^2\big)$ denoting the density function of the normal distribution with mean μ and variance σ^2

→ Determine β and σ^2 such that (log) likelihood function is maximized

Maximum likelihood - algorithms

Expectation Maximization (EM) [Dempster et al., Stewart]

- uses the complete-data likelihood function (assuming precisely known values) to maximize the observed-data likelihood function (based on interval observations).
- Own implementation in R

Survival regression [Kalbfleisch & Prantice], [Meeker & Escobar]

- Used function ‚survreg‘ from R package ‚survival‘ to fit parametric survival regression model (more precisely common slope distinct intercept model)
- ‚survreg‘ is based on likelihood of interval data as well, and uses Newton-Raphson algorithm to maximize the likelihood and determine parameter estimates

Tobit regression [Tobin]

- Special case of survival regression to fit models when the dependent variable is either left- or right-censored (e.g., results <LOQ, but other results have sufficient resolution)
- R package AER provides function ‚tobit‘ which makes use of ‚survreg‘

Maximimum likelihood – adaptions

Residual standard deviation / scale parameter:

- General issue with maximum likelihood: estimate of residual SD is biased because sum of squared deviations is divided by number of observations
- Can make „the usual“ bias correction (corresponding to ordinary least squares estimate) by dividing by number of observations minus number of model parameters (excl. scale parameter)

→ approach/computations using bias correction indicated by „1“, e.g., „EM1“ or „survreg1“

Confidence intervals:

- Maximum likelihood theory usually uses normal distribution for confidence intervals
- could use t-distribution with degrees of freedom equal to number of observations minus number of model parameters (excl. scale parameter)

→ „n“ indicates normal distribution is used, „t“ indicates t-distribution is used, e.g., „EM1t“, „survregn“

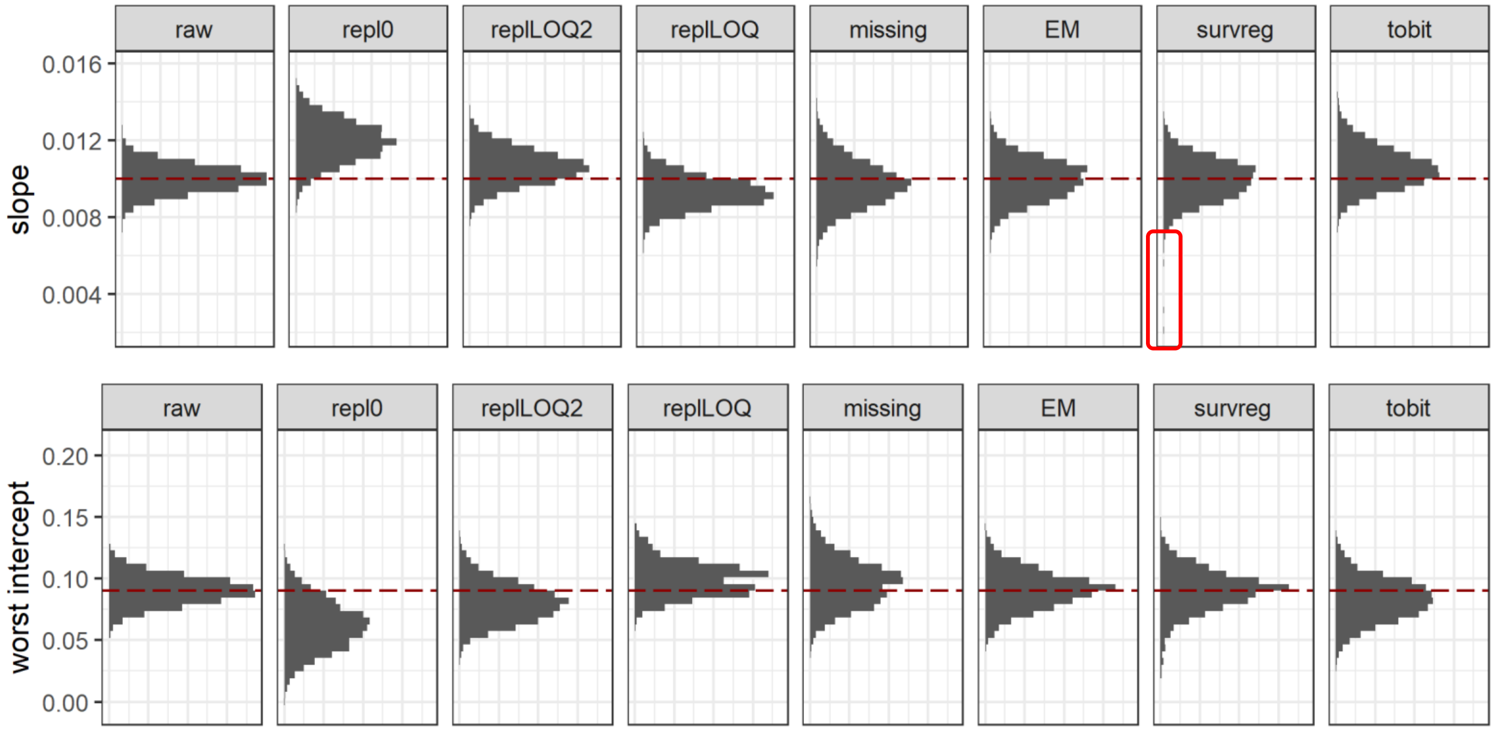
Simulation study – setup

Conduction of a simulation study to obtain general statement about approaches

- 10000 stability data sets were generated using the following inputs:
  - 3 batches with 1 measurement at time points 0, 3, 6, 9, 12, 18, 24 months each
  - Slope: 0.01 %/month for all 3 batches
  - Intercepts: 0.05 % (Batch A), 0.07 % (Batch B), 0.09 % (Batch C)
  - Residual errors randomly drawn from normal distribution with mean 0 and residual SD 0.025 %
  - LOQ 0.095 %, results above LOQ rounded to 1 digit
- For each data set, before-mentioned approaches were used to fit a common slope distinct intercept model
- Histograms and statistics for parameter estimates and shelf lives were plotted/determined over all 10000 data sets
- Shelf life is determined based on upper specification limit of 0.5% → true shelf life is (USL – worst intercept)/slope = (0.5 – 0.09)/0.01 = 41 months

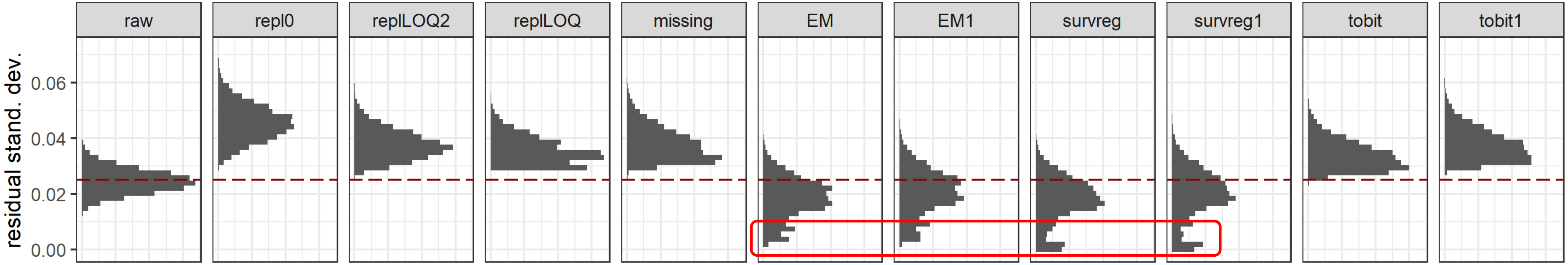
Simulation study – parameter estimates

- Using rounded data together with replacement approaches yields biased parameter estimates (smallest bias LOQ/2)
- treating <LOQ as missing also yields (small) bias
- Same for tobit
- EM and survreg yield unbiased results even though variability is hig-her as with raw data and there are a few outliers with survreg



		raw	repl0	replLOQ2	replLOQ	missing	EM	survreg	tobit
Slope	Mean	0.010003	0.011916	0.010603	0.009290	0.009727	0.009939	0.009931	0.010559
	Median	0.010001	0.011942	0.010585	0.009263	0.009700	0.009965	0.009955	0.010487
	SD	0.000690	0.001072	0.000889	0.000804	0.001197	0.001021	0.001042	0.001047
Worst intercept	Mean	0.090350	0.061848	0.080415	0.099184	0.096034	0.090916	0.089961	0.082541
	Median	0.090298	0.061505	0.080019	0.098973	0.096175	0.091124	0.090912	0.082723
	SD	0.011311	0.019870	0.015499	0.013231	0.019136	0.015250	0.016821	0.016644

Simulation study – residual SD estimate 1/2



	raw	repl0	replLOQ2	replLOQ	missing	EM	EM1	survreg	survreg1	tobit	tobit1
Mean	0.024602	0.045489	0.037612	0.035614	0.037771	0.018316	0.020358	0.017669	0.019638	0.033747	0.037508
Median	0.024471	0.045320	0.037149	0.034898	0.036838	0.018603	0.020676	0.018219	0.020250	0.033065	0.036750

- replacement approaches, treating <LOQ values as missing and tobit: residual SD is clearly over-estimated
- EM and survreg: residual SD is rather under-estimated; and there seems to be a group of runs for which residual SD is clearly under-estimated (close to 0)

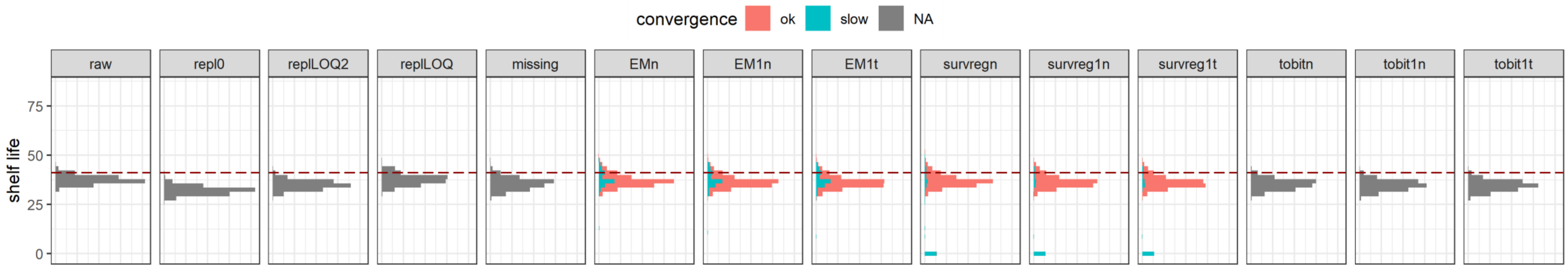
Simulation study – residual SD estimate 2/2

	raw	repl0	replLOQ2	replLOQ	missing	EM	EM1	survreg	survreg1	tobit	tobit1
Mean	0.024602	0.045489	0.037612	0.035614	0.037771	0.020323*	0.022588*	0.019531*	0.021707*	0.033747	0.037508
Median	0.024471	0.045320	0.037149	0.034898	0.036838	0.019692*	0.021887*	0.019228*	0.021370*	0.033065	0.036750

* only computed from data sets with fast convergence (red)

- For (almost all) of the runs for which EM and survreg clearly underestimate the residual SD, the convergence was slow or there was a perfect fit (all predicted values fall inside the intervals used as input data)

Simulation study – shelf life



	raw	repl0	replLOQ2	replLOQ	missing	EMn	EM1n	EM1t	survregn	survreg1n	survreg1t	tobitn	tobit1n	tobit1t
Mean	37.4079	32.1052	34.8472	37.6284	35.4850	36.6269*	36.1702*	35.9132*	36.7413*	36.2906*	36.0375*	35.2140	34.7787	34.5345
Median	37.3000	32.0000	34.9000	37.6000	35.6000	36.3000*	35.9000*	35.6000*	36.4000*	36.0000*	35.7000*	35.3000	34.9000	34.6000
95% quantile	40.9000	35.5000	38.5000	42.4000	40.3000	41.9000*	41.3000*	40.9000*	42.3000*	41.8000*	41.4400*	39.5000	38.9000	38.6000

* only computed from data sets with fast convergence (red)

- To assess appropriateness with respect to shelf life, 95% quantile of simulated shelf lives must be compared to true shelf life of 41 months
  (in 95% of all cases true regression line is covered by 95% CL meaning that CL is wider than regression line → in 95% of all cases intersection of 95% CL with spec is earlier than the intersection of true reg line with spec limit)
- repl0, replLOQ2: too short shelf lives; replLOQ: too long shelf lives; missing: closer to true shelf life
- Tobit (used with rounded data): too short shelf lives
- survreg, EM: work best with bias correction and t-distribution, 95% quantile close to true shelf life when looking only at the data sets with fast convergence

Conclusions

- Using rounded data and replacement approaches not appropriate, missing seems to be the least bad approach, but depends on proportion of <LOQ values
- Tobit with rounded data not appropriate
- EM and survreg work quite well for most cases, BUT
  - be careful regarding convergence and perfect fit
  - Adaptions should be applied:
    - bias correction for residual SD
    - quantile of t-distribution should be used for computation of confidence limits

Overall conclusion: use raw/unrounded data whenever possible, in particular for results above LOQ!

References

Dempster, A.P., Laird, N.M., Rubin D.B., Maximum Likelihood from Incomplete Data via the EM algorithm, Journal of the Royal Statistical Society, Series B (Methodological), 39:1 (1977), 1–38

Stewart, M.B., On least squares estimation when the dependent variable is grouped, The Review of Economic Studies, 50:4 (1983), 737–753

Kalbfleisch, J. D. and Prentice, R. L., The statistical analysis of failure time data, Wiley, 2002

Meeker, W.Q. and Escobar L.A., Statistical methods for reliability data, Wiley, 1998

Tobin, J., Estimation of Relationships for Limited Dependent Variables, Econometrica, 26 (1958), 24–36

Bernhard Schmelzer
bernhard.schmelzer@novartis.com

Thank you

Expectation maximization (EM) algorithm 1/2

General, iterative procedure developed by [Dempster et al. 1977] which uses the complete-data likelihood function (assuming precisely known values) to maximize the observed-data likelihood function (based on interval observations). Its application to linear models was discussed in [Stewart, 1983].

Each iteration consists of 2 steps:

- E-step: determine conditional expectation of complete-data log-likelihood given the interval data using β and σ^2 from previous iteration
- M-step: find β and σ^2 such that this conditional expectation is maximized

Expectation maximization (EM) algorithm 2/2

If $\widehat{\beta}_{(k)}, \widehat{\sigma}^2_{(k)}$ denote the estimates obtained by the k-th iteration, then the estimates of iteration (k+1) are obtained by the following formulas

$$\widehat{\beta}_{(k+1)} = (X^\top X)^{-1} X^\top \mathrm{E}_{(k)}(Y_i|\mathbb{Y}_i) \qquad \widehat{\sigma}^2_{(k+1)} = \frac{1}{n}\Big(\big(\sum_{i=1}^{n} \mathrm{E}_{(k)}(Y_i^2|\mathbb{Y}_i)\big) - \big(\mathrm{E}_{(k)}(Y|\mathbb{Y})\big)^\top X (X^\top X)^{-1} X^\top \mathrm{E}_{(k)}(Y|\mathbb{Y}) \Big)$$

Hence the formulas are basically the same as for classical maximum likelihood but instead of the observed data conditional expectations of moments based on parameter estimates from k-th iteration are used:

$$\mathrm{E}_{(k)}(Y_i^m|\mathbb{Y}_i) = \int_{\underline{y}_i}^{\overline{y}_i} y_i^m \, g\big(y_i \big| X_{i*}\widehat{\beta}_{(k)}, \widehat{\sigma}^2_{(k)}\big) \, \mathrm{d}y_i$$

Iterations are repeated until desired precision in estimates is achieved.

Interval midpoints were used for obtaining starting values, choice might be crucial.

Implementation by own R code for performing computations.

Survival and Tobit regression

Survival regression

- Used function ‚survreg' from R package ‚survival' to fit parametric survival regression model (more precisely common slope distinct intercept model)
- ‚survreg' is based on likelihood of interval data as well, and uses Newton-Raphson algorithm to maximize the likelihood and determine parameter estimates
- References: [Kalbfleisch & Prantice], [Meeker & Escobar]
- used to compare to EM implementation

Tobit regression

- Special case of survival regression to fit models when the dependent variable is either left- or right-censored (e.g., results <LOQ, but other results have sufficient resolution)
- R package AER provides function ‚tobit' which makes use of ‚survreg'
- Reference: [Tobin]

Stability data example – results of interval approaches

Approach	Slope	Worst intercept	Resid. SD	Shelf life
Unrounded (raw)	0.009956	0.088173	0.026566	37.4
EM normal (EMn)	0.009296	0.097603	0.017162	38.0
EM bias corr, normal (EM1n)	0.009296	0.097603	0.019075	37.5
EM bias corr, t-dist (EM1t)	0.009296	0.097603	0.019075	37.2
Survreg normal (survregn)	0.009296	0.097603	0.017162	38.0
Survreg bias corr, normal (survreg1n)	0.009296	0.097603	0.019074	37.5
Survreg bias corr, t-dist (survreg1t)	0.009296	0.097603	0.019074	37.2
Tobit normal (tobitn)	0.010136	0.081450	0.028277	37.0
Tobit df.corr, normal (tobit1n)	0.010136	0.081450	0.031428	36.6
Tobit df.corr, t-dist (tobit1t)	0.010136	0.081450	0.031428	36.4

EM and survreg results are very similar:

- Slope (slightly) under-estimated
- Intercept (slightly) over-estimated
- Residual SD under-estimated
- Shelf lives close to SL of unrounded data when using bias correction

Tobit:

- Slope estimate close to the one based on unrounded data
- Intercept under-estimated
- Residual SD over-estimated
- Shelf lives rather too short

Stability data example – true values and model

The stability data from the previous slides was generated using the following inputs:

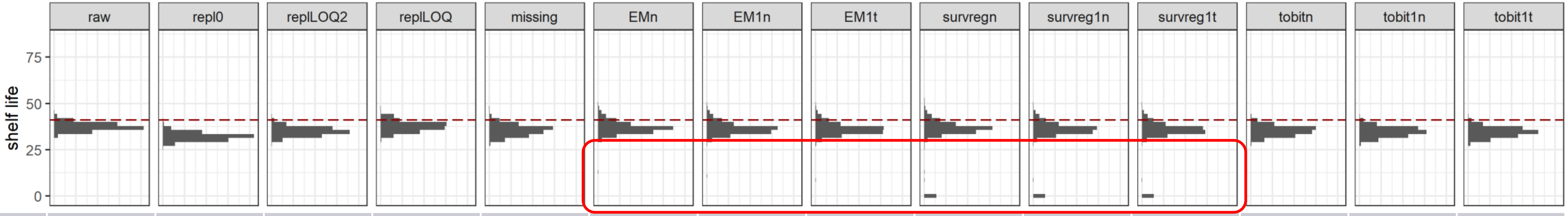
- 3 batches with 1 measurement at time points 0, 3, 6, 9, 12, 18, 24 months each
- Slope: 0.01 for all 3 batches
- Intercepts: 0.05 (Batch A), 0.07 (Batch B), 0.09 (Batch C)
- Residual errors randomly generated from normal distribution with mean 0 and standard deviation (SD) 0.025

Throughout this presentation, a fixed effect model assuming common slope for all batches but distinct intercepts (t0 values) for each batch will be considered:

$$y_{ki} = a_k + b \cdot t_{ki} + \epsilon_{ki}$$

k index for batches, i index for measurements, a_k intercept of Batch k, b common slope, y_{ki} quality attribute measurements, t_{ki} time values, ϵ_{ki} measurement error

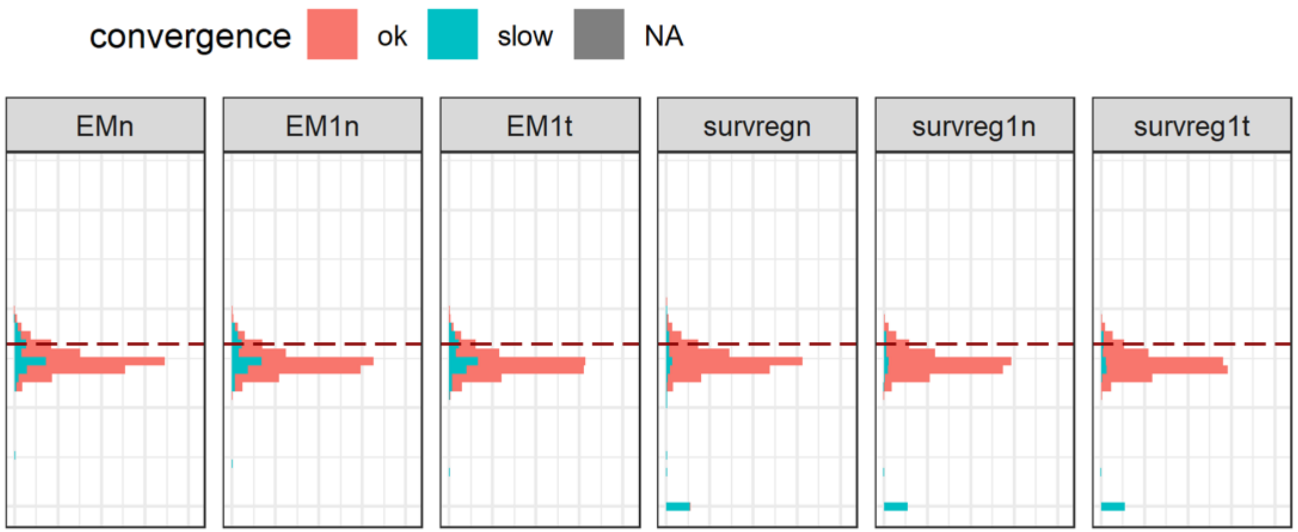
Simulation study – shelf life 1/2



	raw	repl0	replLOQ2	replLOQ	missing	EMn	EM1n	EM1t	survregn	survreg1n	survreg1t	tobitn	tobit1n	tobit1t
Mean	37.4079	32.1052	34.8472	37.6284	35.4850	36.7746	36.3234	36.0681	Inf	Inf	Inf	35.2140	34.7787	34.5345
Median	37.3000	32.0000	34.9000	37.6000	35.6000	36.5000	36.0000	35.8000	36.3	35.9	35.6	35.3000	34.9000	34.6000
95% quantile	40.9000	35.5000	38.5000	42.4000	40.3000	42.6000	42.1000	41.7000	42.6	42.0	41.7	39.5000	38.9000	38.6000

- To assess appropriateness with respect to shelf life, 95% quantile of simulated shelf lives must be compared to true shelf life of 41 months
  (in 95% of all cases true regression line is covered by 95% CL meaning that CL is wider than regression line → in 95% of all cases intersection of 95% CL with spec is earlier than the intersection of true reg line with spec limit)
- repl0, replLOQ2: too short shelf lives; replLOQ: too long shelf lives; missing: closer to true shelf life
- Tobit (used with rounded data): too short shelf lives
- survreg, EM: work best with bias correction and t-distribution, 95% quantile close to true shelf life
  - some outliers resulting in too short shelf lives
  - survreg: for 2 out of 10000 runs slope and residual SD estimate are so small that shelf life >100 months → displayed as “Inf”

Simulation study – shelf life 2/2



	EMn	EM1n	EM1t	survregn	survreg1n	survreg1t
Mean	36.6269*	36.1702*	35.9132*	36.7413*	36.2906*	36.0375*
Median	36.3000*	35.9000*	35.6000*	36.4000*	36.0000*	35.7000*
95% quantile	41.9000*	41.3000*	40.9000*	42.3000*	41.8000*	41.4400*

* only computed from data sets with fast convergence (red)

- Outliers in shelf life histograms come from runs with slow convergence or perfect fit
- Excluding these runs, the 95% quantile of the simulated shelf lives is close to the true shelf life of 41 months when using bias correction and t-distribution with EM and survreg (EM1t, survreg1t)