Process characterization: Correlation structure between two Critical Quality Attributes when there are multiple measurements per condition

Jean-François Michiels |
Associate Director Statistics |
Pharmalex

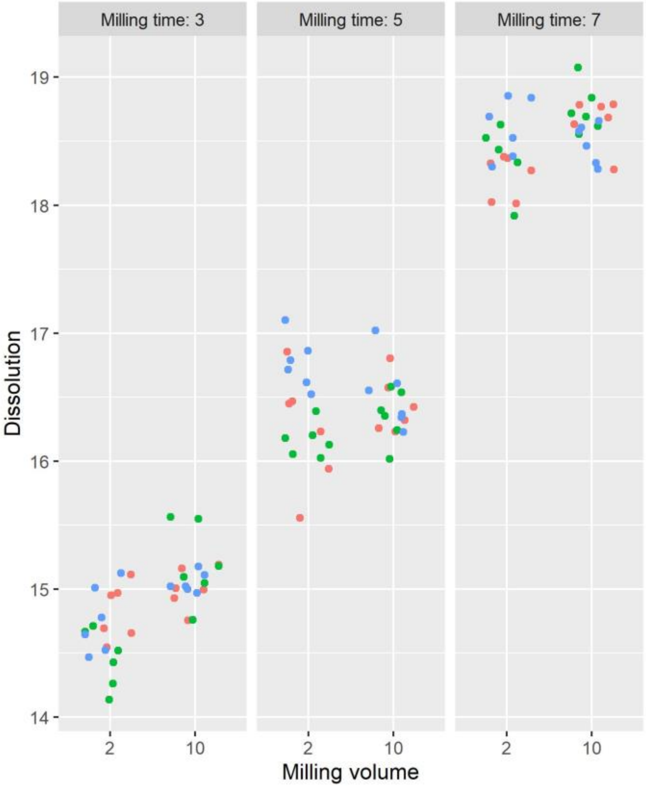
21 Oct 2022

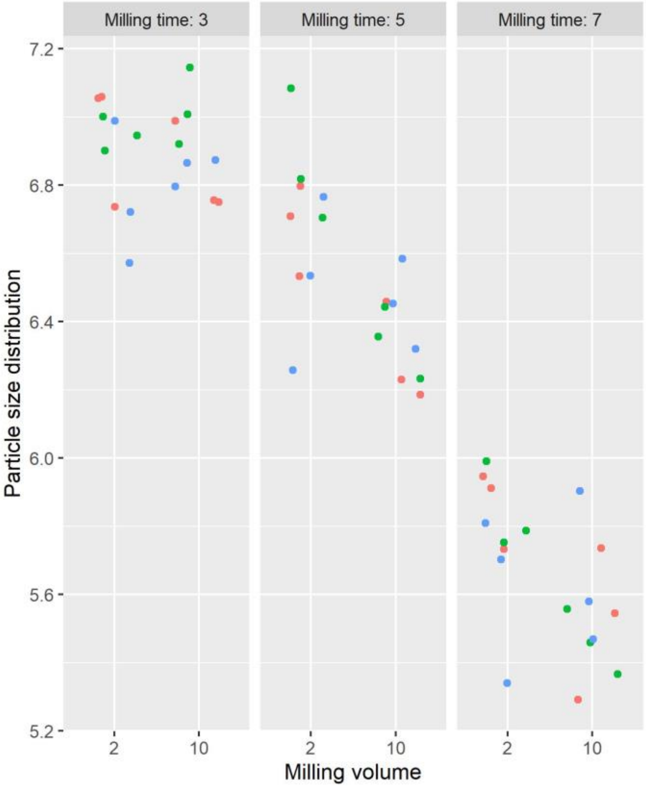
PHARMALEX

Illustrative example

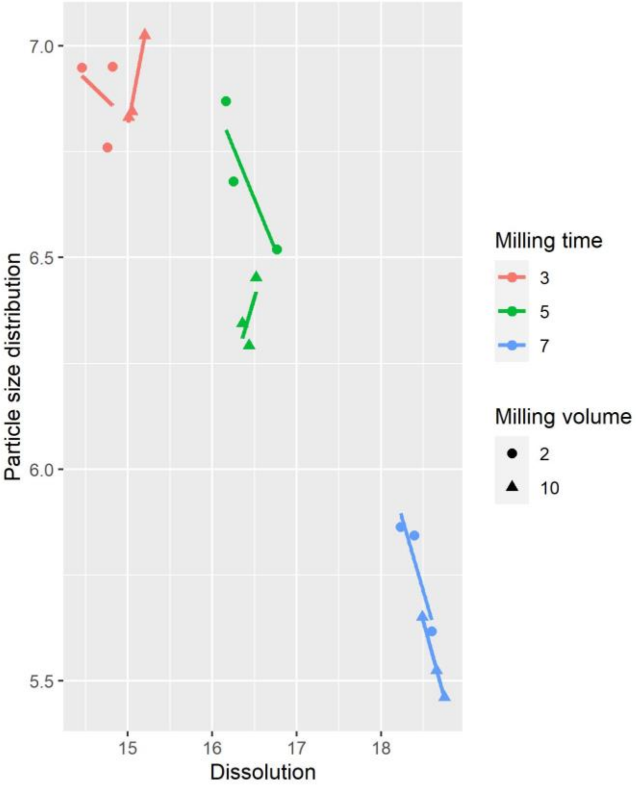
- Characterization of a powder/tablet manufacturing process
- CPPs:
  - milling time (3 levels)
  - milling volume (2 levels)
- CQAs:
  - Particle Size Distribution (PSD) 3 measurement per batch
  - Dissolution 6 measurements per batch
- 3 batches per manufacturing condition
- Model: Main effects and interactions

Data (Batch are colored within each manufacturing condition)





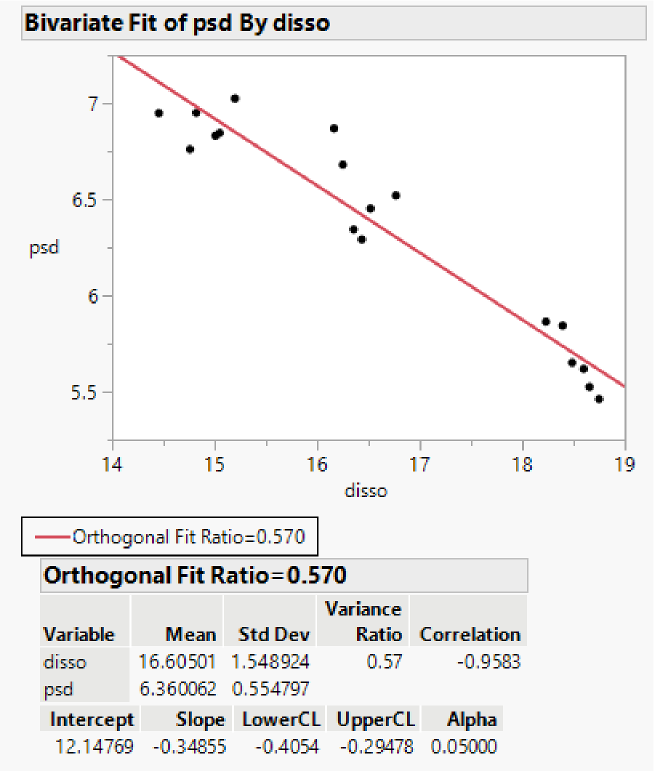
Aggregated data



- Within a single set of manufacturing conditions, a negative correlation is observed between batch
  - If psd is increasing, dissolution is decreasing

Error-in-variables regression

- Using the average of the data (mean of 6 and 3 values respectively for dissolution and psd)
- Variance ratio is estimated from the data
- This approach makes difficult to interpret the effects of CPPs on both CQAs



Bivariate Fit of psd By disso

Orthogonal Fit Ratio=0.570

Variable	Mean	Std Dev	Variance Ratio	Correlation
disso	16.60501	1.548924	0.57	-0.9583
psd	6.360062	0.554797		

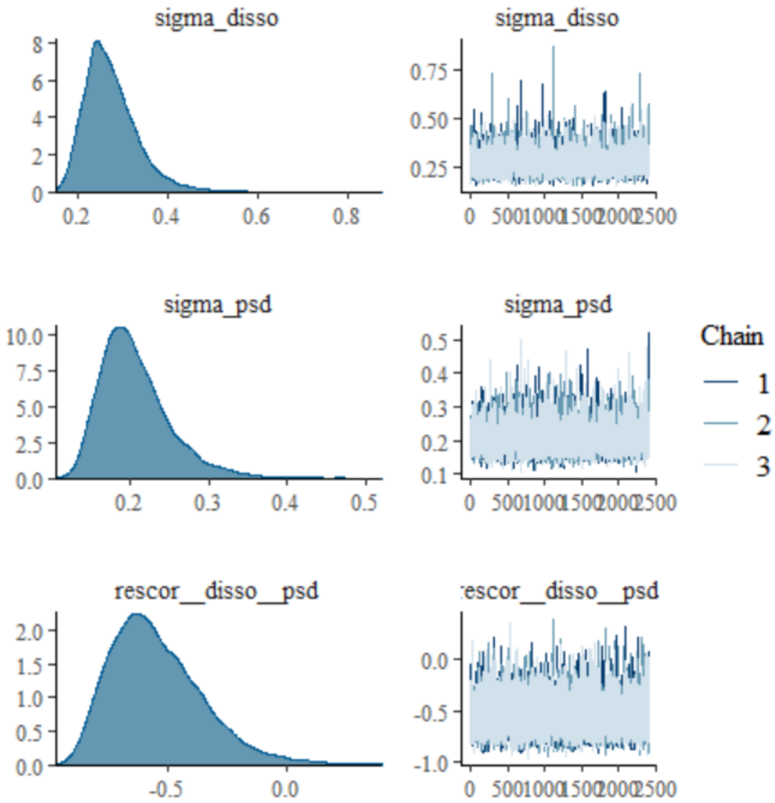
Intercept	Slope	LowerCL	UpperCL	Alpha
12.14769	-0.34855	-0.4054	-0.29478	0.05000

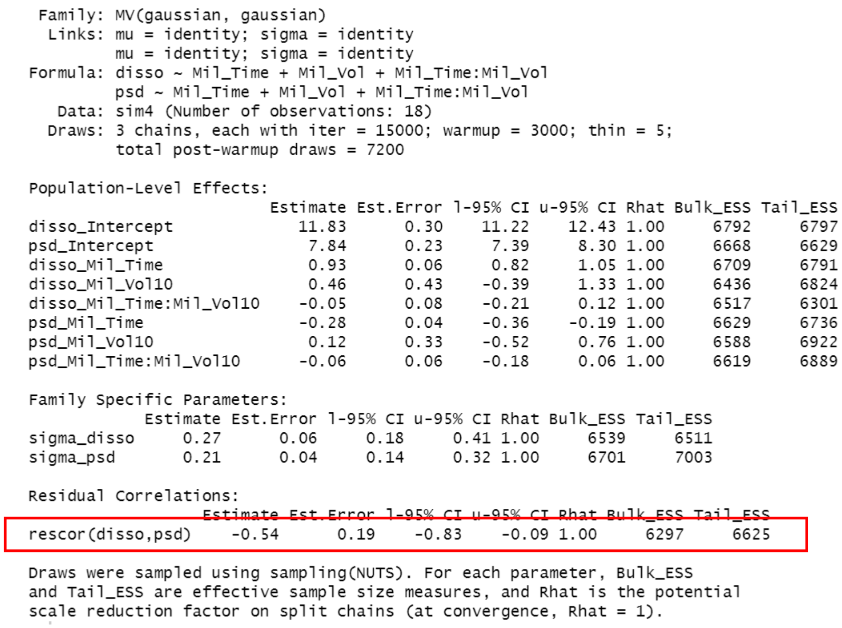
Bi-response model on average data

- Based on average data
- A single value per batch
- Residual error includes mostly batch-to-batch variability
- The correlation between residuals is estimated

```
form <- bf(mvbind(disso,psd)~Mil_Time+Mil_Vol+Mil_Time:Mil_Vol) + set_rescor(TRUE)
fit2 <- brm(form,  data=sim4,warmup=3000,thin=5,iter=15000,chains=3)
```

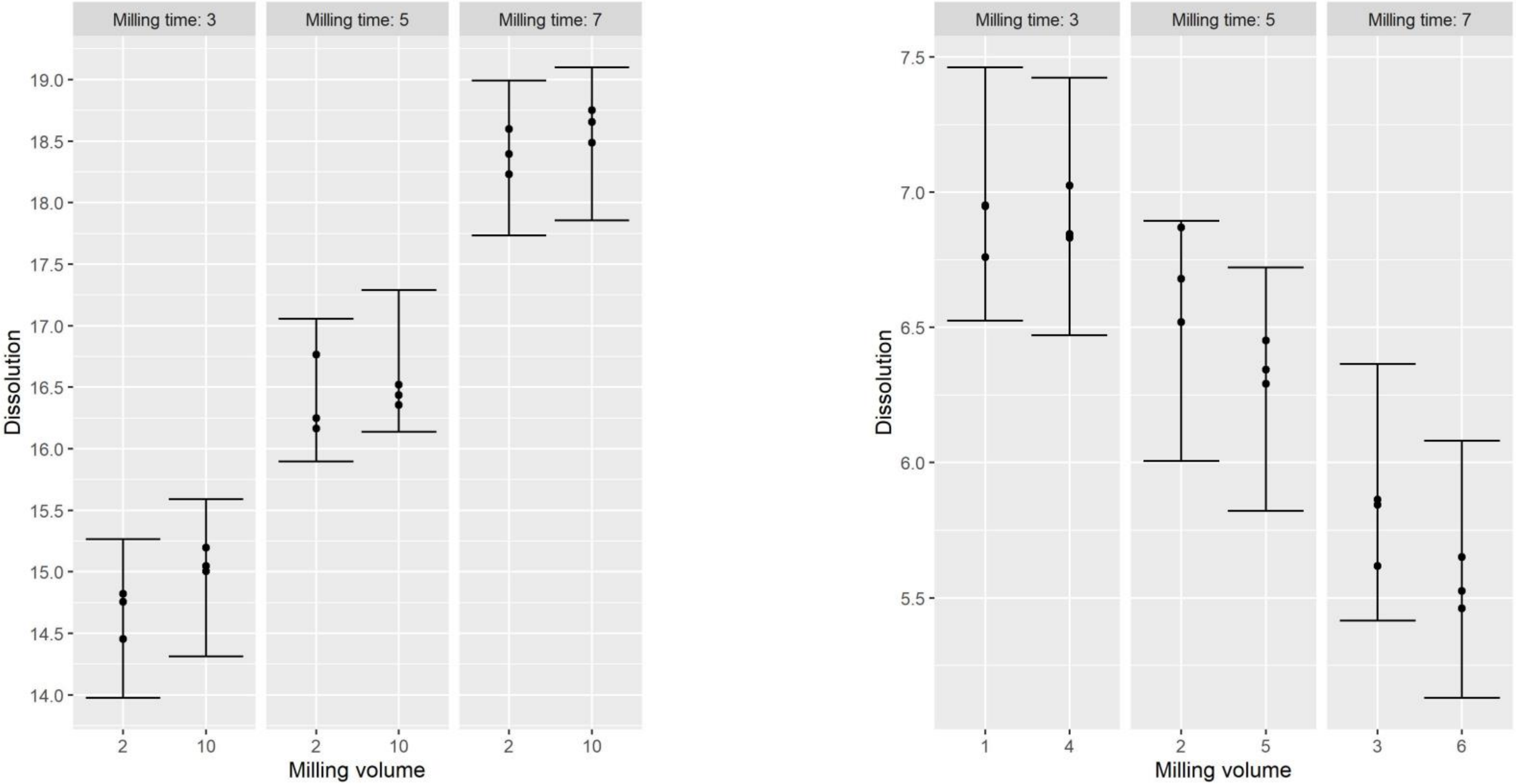
Results of the average model





```
 Family: MV(gaussian, gaussian)
  Links: mu = identity; sigma = identity
         mu = identity; sigma = identity
Formula: disso ~ Mil_Time + Mil_Vol + Mil_Time:Mil_Vol
         psd ~ Mil_Time + Mil_Vol + Mil_Time:Mil_Vol
   Data: sim4 (Number of observations: 18)
  Draws: 3 chains, each with iter = 15000; warmup = 3000; thin = 5;
         total post-warmup draws = 7200

Population-Level Effects:
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
disso_Intercept             11.83      0.30    11.22    12.43 1.00     6792     6797
psd_Intercept                7.84      0.23     7.39     8.30 1.00     6668     6629
disso_Mil_Time               0.93      0.06     0.82     1.05 1.00     6709     6791
disso_Mil_Vol10              0.46      0.43    -0.39     1.33 1.00     6436     6824
disso_Mil_Time:Mil_Vol10    -0.05      0.08    -0.21     0.12 1.00     6517     6301
psd_Mil_Time                -0.28      0.04    -0.36    -0.19 1.00     6629     6736
psd_Mil_Vol10                0.12      0.33    -0.52     0.76 1.00     6588     6922
psd_Mil_Time:Mil_Vol10      -0.06      0.06    -0.18     0.06 1.00     6619     6889

Family Specific Parameters:
            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma_disso     0.27      0.06     0.18     0.41 1.00     6539     6511
sigma_psd       0.21      0.04     0.14     0.32 1.00     6701     7003

Residual Correlations:
                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
rescor(disso,psd)    -0.54      0.19    -0.83    -0.09 1.00     6297     6625

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

95% PI using the average model



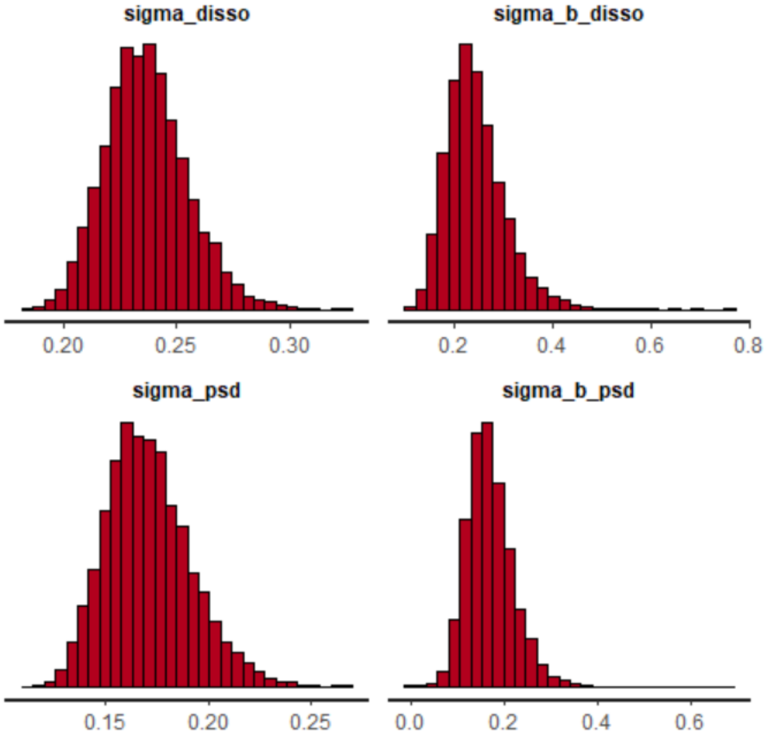
Bi-response model with individual data

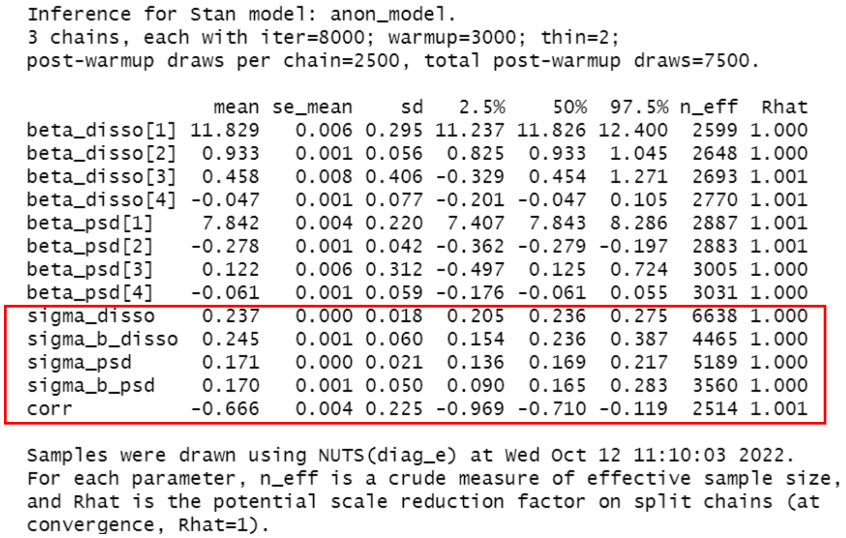
- Performed in Stan (rstan v2.26)
- Several value per batch → random batch effect is possible
- Correlation is set on the batch variability between the two responses

```
transformed parameters{
     matrix[2,2] sigma_matrix;
     sigma_matrix[1,1] = sigma_b_disso*sigma_b_disso;
     sigma_matrix[2,2] = sigma_b_psd*sigma_b_psd;
     sigma_matrix[1,2] = corr*sigma_b_disso*sigma_b_psd;
     sigma_matrix[2,1] = corr*sigma_b_disso*sigma_b_psd;
 }
```

```
model {
 // vector definitions

 for(j in 1:N){
   for(i in 1:6){
     Y[j , i] ~ normal(beta_disso[1] +
                       beta_disso[2]*Mil_Time[j] +
                       beta_disso[3]*Mil_Vol10[j]+
                       beta_disso[4]*Mil_Time_Mil_Vol10[j]+
                       a_batch[Batch[j],1],
                      sigma_disso);
   }

   for(i in 7:9){
     // similar to the psd model
   }

 }

 a_batch ~ multi_normal(a_0,sigma_matrix);
 //other prior definitions
}
```

Results of the model with individual data

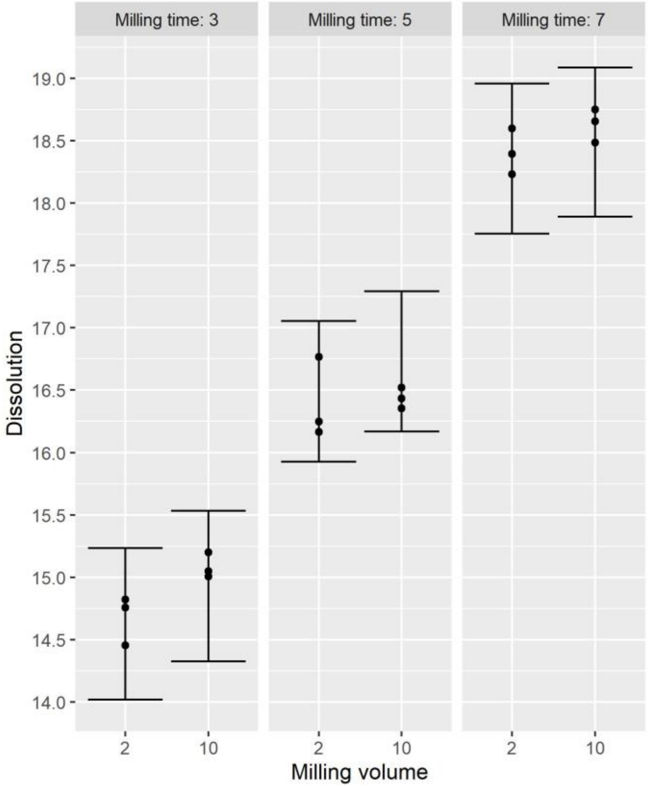


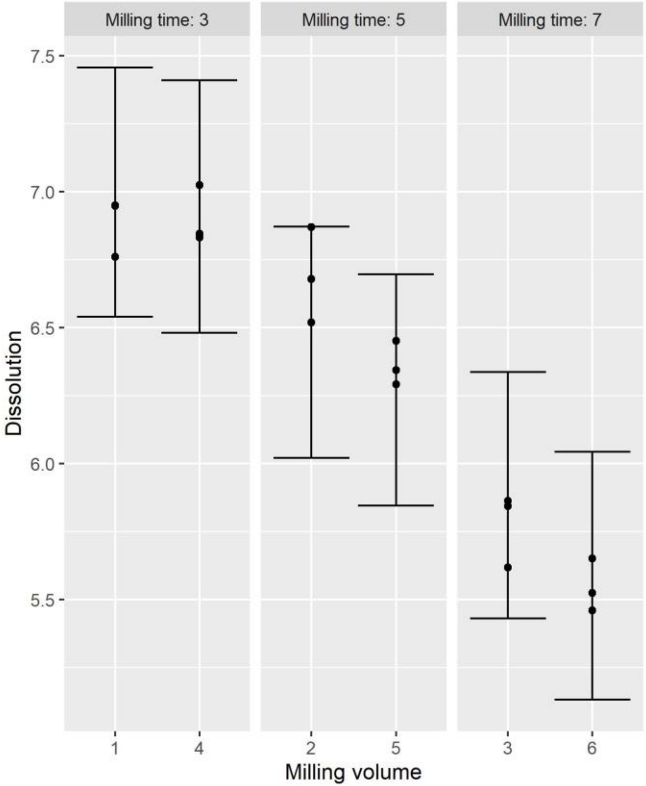


```
Inference for Stan model: anon_model.
3 chains, each with iter=8000; warmup=3000; thin=2;
post-warmup draws per chain=2500, total post-warmup draws=7500.

                 mean se_mean    sd   2.5%    50%  97.5% n_eff  Rhat
beta_disso[1] 11.829   0.006 0.295 11.237 11.826 12.400  2599 1.000
beta_disso[2]  0.933   0.001 0.056  0.825  0.933  1.045  2648 1.000
beta_disso[3]  0.458   0.008 0.406 -0.329  0.454  1.271  2693 1.001
beta_disso[4] -0.047   0.001 0.077 -0.201 -0.047  0.105  2770 1.001
beta_psd[1]    7.842   0.004 0.220  7.407  7.843  8.286  2887 1.001
beta_psd[2]   -0.278   0.001 0.042 -0.362 -0.279 -0.197  2883 1.001
beta_psd[3]    0.122   0.006 0.312 -0.497  0.125  0.724  3005 1.000
beta_psd[4]   -0.061   0.001 0.059 -0.176 -0.061  0.055  3031 1.000
sigma_disso    0.237   0.000 0.018  0.205  0.236  0.275  6638 1.000
sigma_b_disso  0.245   0.001 0.060  0.154  0.236  0.387  4465 1.000
sigma_psd      0.171   0.000 0.021  0.136  0.169  0.217  5189 1.000
sigma_b_psd    0.170   0.001 0.050  0.090  0.165  0.283  3560 1.000
corr          -0.666   0.004 0.225 -0.969 -0.710 -0.119  2514 1.001

Samples were drawn using NUTS(diag_e) at Wed Oct 12 11:10:03 2022.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Correlation is better estimated
Total variability from average data is consistent with the sum of the model with individual data
For psd:
Sqrt(0.171^2+0.170^2)=0.241
Average model was equal to 0.21

95% PI using the model with individual data





Note: This is the PI of the average values to make it comparable to the results of the average model

Probabilities of success

- Assuming specs
  - For disso: [15,20]
  - For psd: [6,10]
- Probabilities of success for all 6 disso and 3 psd values to be within specs

Milling Time	Milling Vol	PoS disso	PoS psd
3	2	2.3	100.0
5	2	99.9	90.7
7	2	100.0	14.2
3	10	12.9	99.8
5	10	100.0	70.9
7	10	99.9	1.4

Conclusions

- Average model and model using individual data are consistent
- Advantages of model using individual data (over average model) makes possible
  - To fit a mixed model and obtain individual replicate variance information
  - To obtain better correlation estimates
  - To obtain the posterior predictive distribution of individual data
  - To get the probabilities of success for any result format (e.g. control strategy)