SSD Approach to Quantify Uncertainty from QSAR Estimates of Effects Data: Examples from CADASTER Classes

Tom Aldenberg

RIVM, Bilthoven, NL

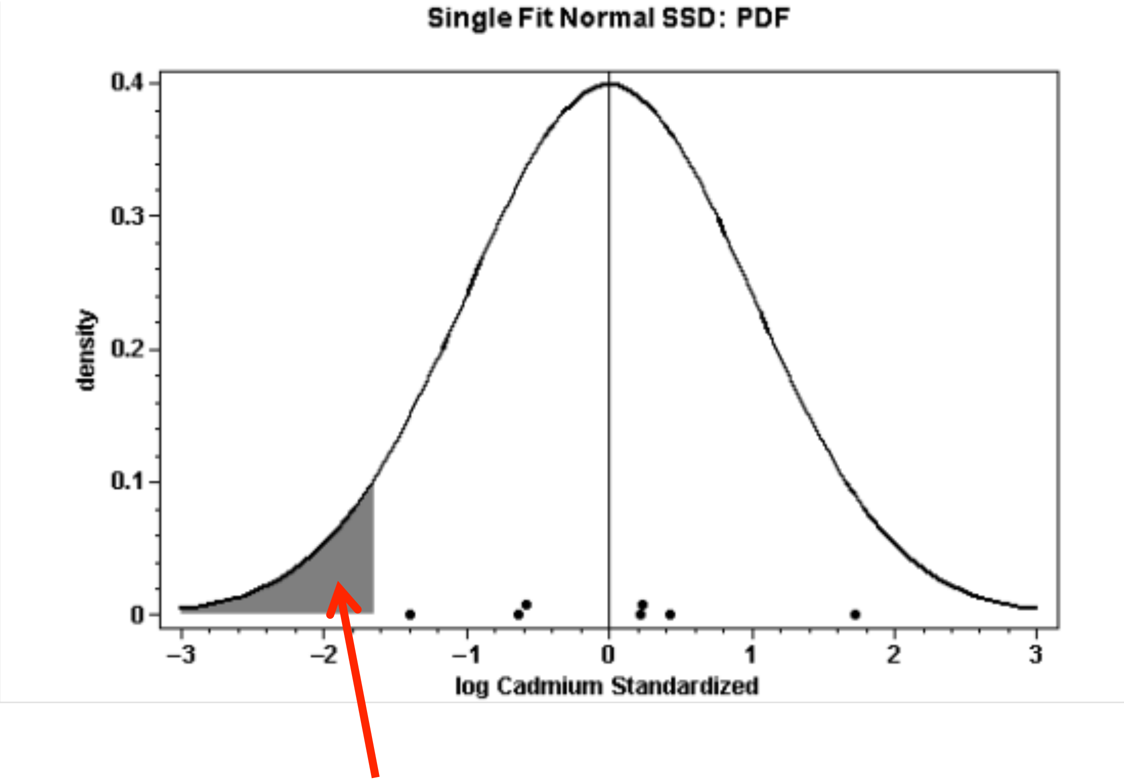
CADASTER Workshop 2012, Oct. 7-9

Contents of the presentation

- Recapitulation of SSD handling, the Bayesian way
- From SSD to QSARs, the Bayesian way
- Importing QSAR uncertainty into an SSD: options
- Summarizing

NOEC or Acute Species Sensitivities: single Normal fit through Mean and Standard deviation of the data

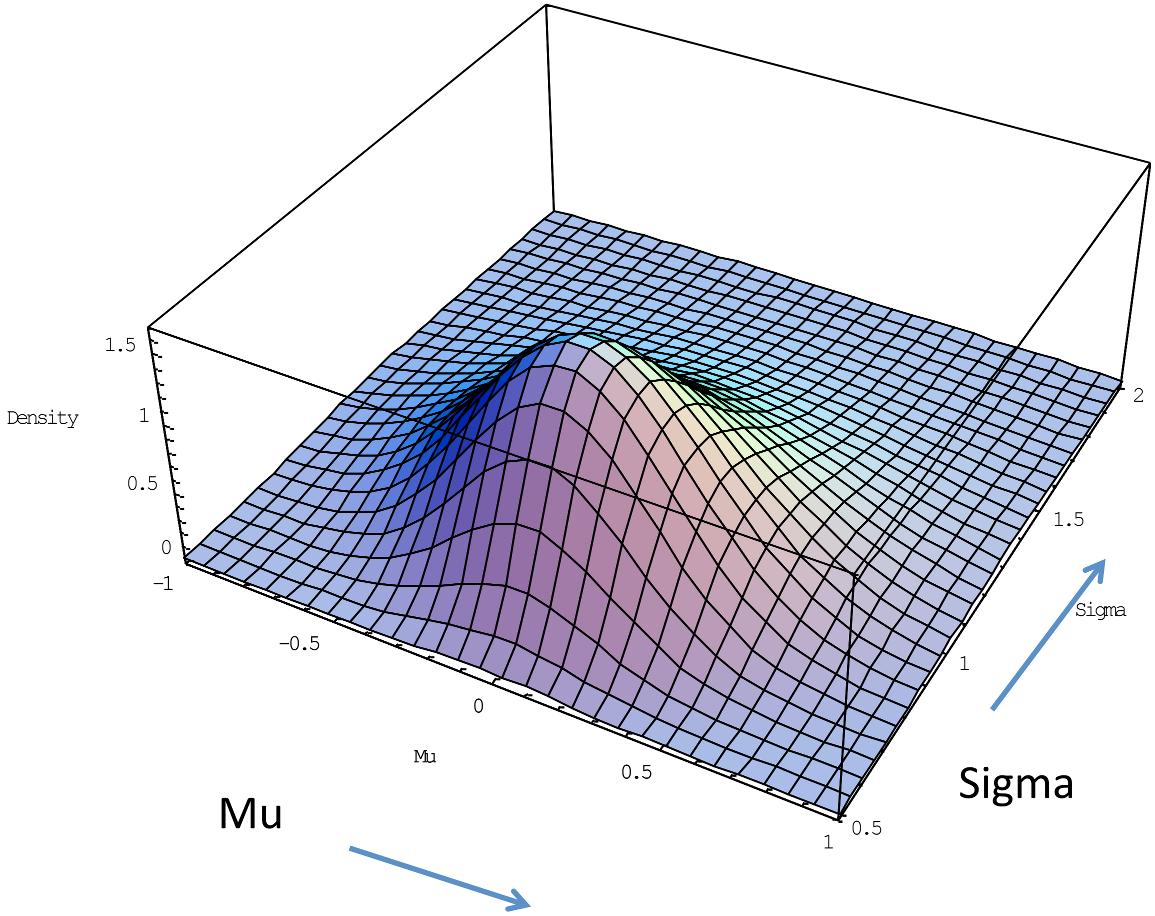
- SSD: distribution of species toxicities (NOECs, EC_{50}s) for a selection of biological species
- Small sample, REACH: minimum 10, preferably 15, spread over 8 taxonomic groups
- Simplest model: Normal distribution, estimated from *sample mean* and *standard deviation*
- Shaded area: Fraction Affected (FA) for 5% of the species
- Hazardous concentration ($\log_{10}$ HC_5), at –1.645*standard deviation below the mean



5%

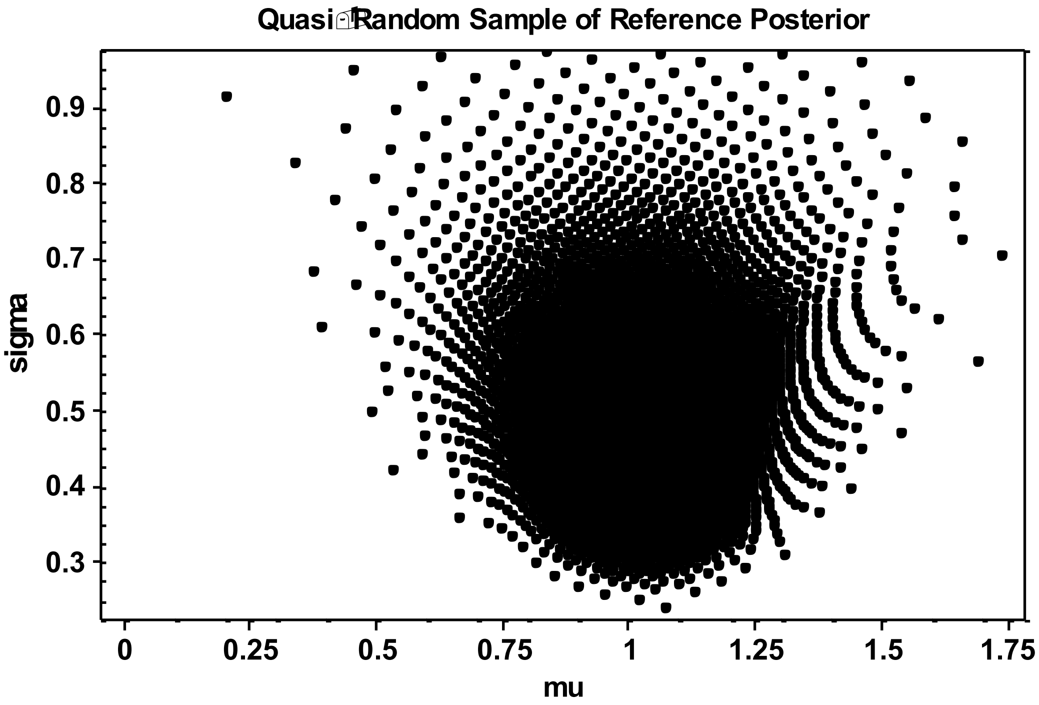
SSD Uncertainty from Bayesian Second-Order Gaussian Fit (1): Posterior Mu and Sigma

- For small samples the Gaussian fit is uncertain
- In Bayesian Statistics the parameters of a model are themselves distributed
- Prior mu and log(sigma) distributions are uniform
- Posterior distribution of mu and sigma uncertainty is a known *bivariate* distribution



SSD Uncertainty from Bayesian Second-Order Gaussian Fit (2): Monte Carlo sample of posterior Mu and Sigma

- Mu-sigma sample is drawn from the posterior distribution
- Each point is a (single curve) Gaussian
- Thus, we have collection of Gaussian curves



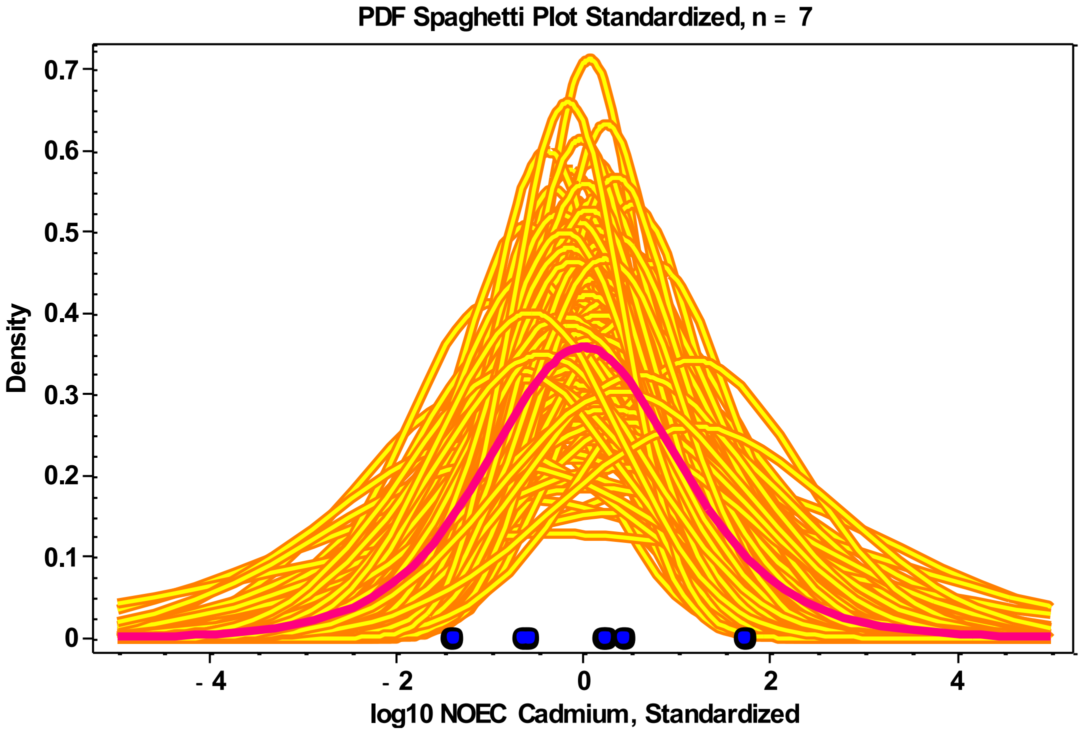
SSD Uncertainty from Bayesian Second-Order Gaussian Fit (3): 'Spaghetti Plot' and Predictive Distribution

- Individual Gaussians form a collection of Gaussian distributions
- The predictive distribution is the mean PDF, that is: average of PDF curves
- The mean PDF is known to be a Student-t, with location, scale, and degrees of freedom:

$$\hat{\alpha} = \overline{x}$$

$$\hat{\beta} = \sqrt{1 + \frac{1}{n}} \cdot s_{n-1}$$

$$\nu = n - 1$$



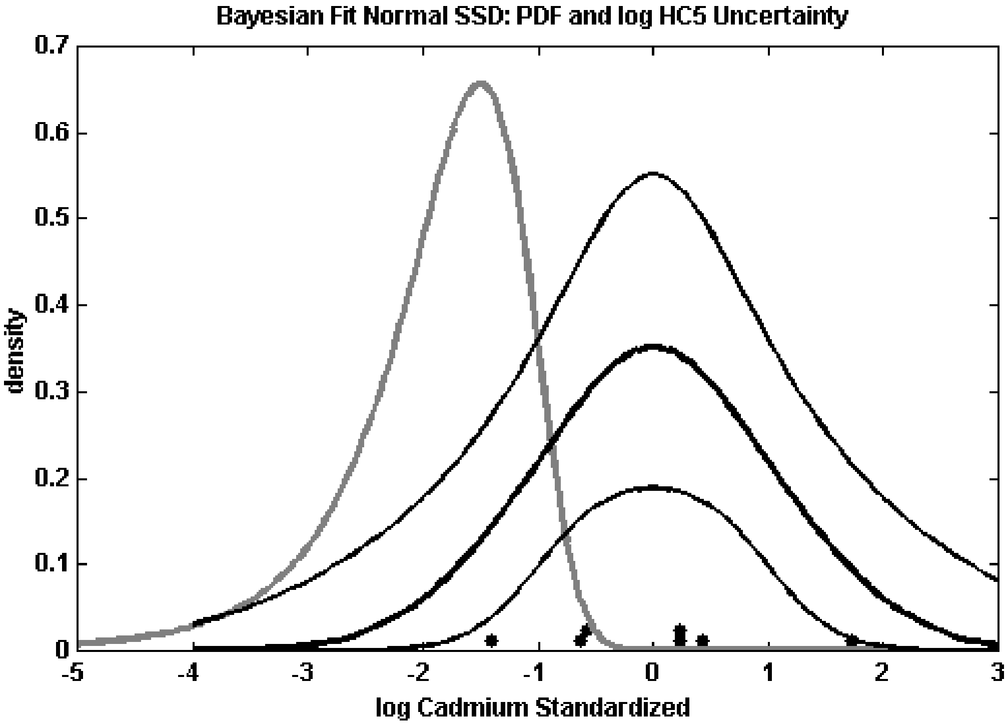
SSD Uncertainty from Bayesian Second-Order Gaussian Fit (4): Percentile Curves and PNEC Uncertainty

- **Percentile Curves (vertical, point-wise percentiles)**
- **$\log_{10}$ HC$_5$ is distributed: uncertainty for the small sample size**
- **Percentiles of the $\log_{10}$ HC$_5$ distribution are estimated from mean and std of the data:**

$$\log_{10} HC_5 = \bar{x} - k_n \cdot s$$

- **The k_n are tabulated as *Extrapolation Constants*:**

$$\log_{10} {HC_5}^{\bar{x}=0, s=1} = -k_n$$

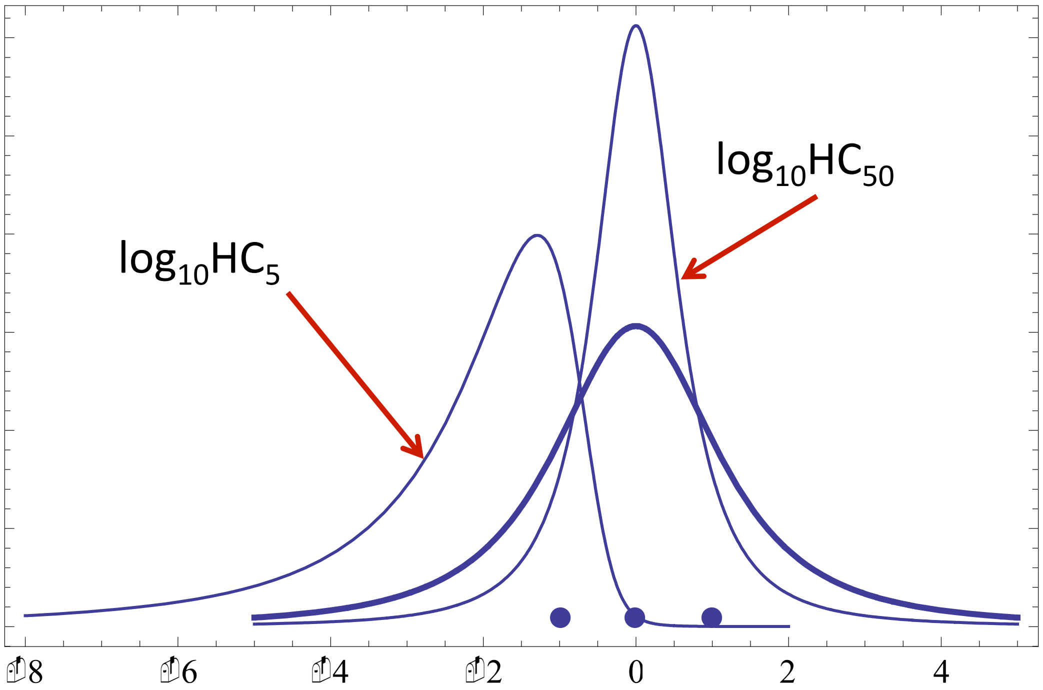


Posterior ('Predictive') Distributions: $\log_{10}HC_5$ and $\log_{10}HC_{50}$ (thin lines). The case n = 3

- *Any quantity* derivable from the parameters (mu, sigma) is distributed (uncertain)
- Mean (Expected value): $\log_{10}HC_{50}$ = μ
- 5th Percentile: $\log_{10}HC_5$ =

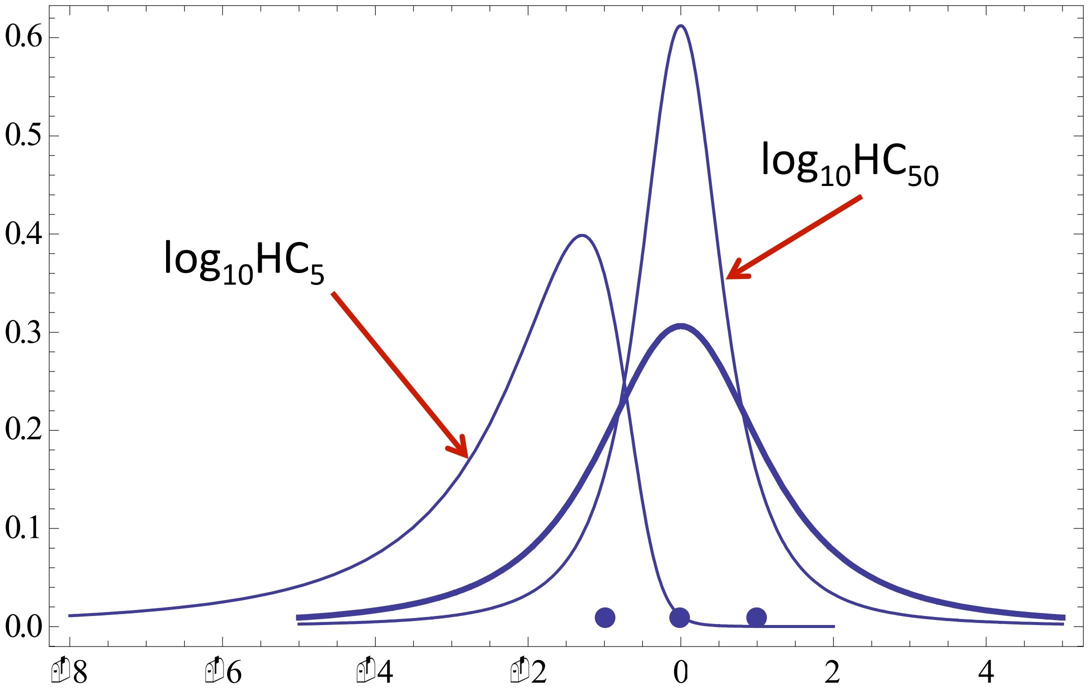
$$\mu - 1.645 \cdot \sigma$$

- If the parameters were known, these were point estimates
- One may summarize their uncertainty with other point estimates, e.g. mean or percentiles



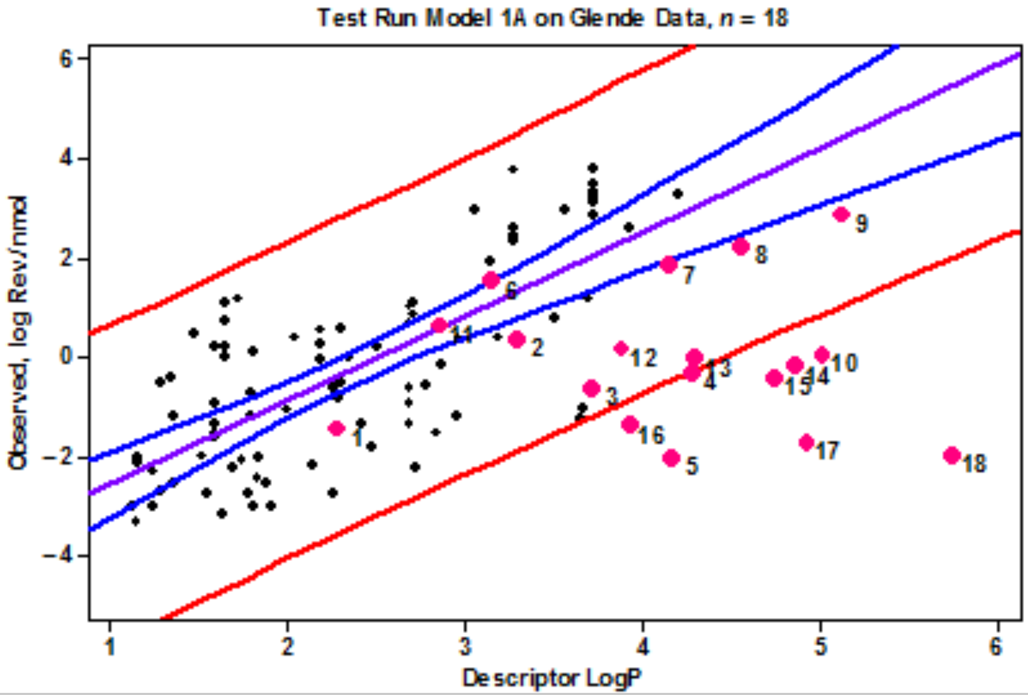
Mean PDF calculated vertically (thick line): probability for a new data point

- Bayesians often indicate the probability of new data as THE predictive distribution
- It is, since the PDF is also a function of mu and sigma
- However, as it is just one of many, it is not unique
- For this model, the Bayesian values exactly match the confidence limits of the mean (–2.48, +2.48) and the predictive limits (–4.97, +4.97)
- Note that the median $\log_{10}HC_5$, often used for the PNEC (!), equals just –1.94, i.e. (well) within both sets of limits



QSAR Regressions (MLR) with Normal Error: Same Predictive Possibilities

- Both types of symmetric predictive limits (expected value, vs. new data) transfer naturally to QSAR Regressions: linear models with Normal error
- Which one to use? This depends on the purpose, obviously
- Want to validate test data? Use the probability of new data
- Want to assess model prediction uncertainty? Use the uncertainty of the expected value
- Want to estimate a quantile at some input? Use the uncertainty of the quantile
- And so on



Question: could this be a standard option in the QSAR Toolbox?

Textbook Regression (classical) is almost all you need for the Bayesian uncertainty as well

- **OLS Fit:**

$$\hat{y} = X \cdot \hat{\beta}$$

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

- **MSE (*k* predictors), estimate of sigma^2:**

$$s^2 = \frac{1}{n-(1+k)} \left(y - X \cdot \hat{\beta}\right)^T \left(y - X \cdot \hat{\beta}\right)$$

- **Student-t scale for the model estimate uncertainty (new substance with predictor vector *x_0*):**

$$s \cdot \sqrt{x_0^T \cdot \left(X^T X\right)^{-1} \cdot x_0}$$

- **Student-t scale for the new data uncertainty (same new substance):**

$$s \cdot \sqrt{1 + x_0^T \cdot \left(X^T X\right)^{-1} \cdot x_0}$$

Combining QSAR Predictions into an SSD

- Experimental Species Data is SSDs are usually taken as fixed
- If the SSD is a distribution of Species Means (which seems to be the idea in the experimental data handling), then substantially uncertain (predictive) means would not add to the SSD variance, but REDUCE IT: Variance Components Argument
- We don't go that route, since we do not want highly uncertain QSARs to diminish (or even destroy) species toxicity variability
- We propose to evaluate the model estimate uncertainty of the QSARs (as ***model uncertainty***), take the appropriate Student-t MC samples from each QSAR for a new substance and put these in SSDs evaluated as full predictive species data uncertainty (as ***data uncertainty***) as the predictive data Student-t

QSAR Model Estimate Uncertainties adds Manageable Conservatism to the SSD

- Although the QSAR model prediction uncertainty adds conservatism (i.e. variance) to the SSD, we do accept that as healthy, knowing that more and/or better data in the QSARs will reduce this conservatism, until (for 'ideal' QSARs), the SSD converges to the SSD of species point estimates
- Experimental points, either taken as fixed data (classically), or in the same vain applying a model (e.g. dose-response) to get a model estimate uncertainty for these data can be combined with the QSAR-based points (uncertainties)
- If we would have used the full predictive data uncertainty in the QSARs, then even huge increases of training data points would not reduce this conservatism, since the data uncertainty predictive limits are not reduced by more points in the QSAR, while the model uncertainty estimates get better with more data

Summarizing adding QSARs to SSDs

- Recapitulation of SSD handling, the Bayesian way
  - Mu, Sigma Posterior Distribution
  - Spaghetti Plot and SSD data predictive distribution
  - $Log_{10}HC_5$, $Log_{10}HC_{50}$ uncertainty, and New Data uncertainty
  - Which Predictive Uncertainty? Depends on the purpose!
  - Median $Log_{10}HC_5$ (PNEC) compared to the two 'universal' predictive distributions
- From SSD to QSARs, the Bayesian way
  - The SSD uncertainties are a special case, mathematically, of (QSAR) regression including predictors
  - Bayesian version is almost textbook version
- Importing QSAR uncertainty into an SSD: options
  - Adding QSAR uncertainty to the SSD in a mildly conservative way
  - More QSAR data should reduce this conservativeness, not confirm it