Deterministic vs. Stochastic Modeling of Uncertainty

1 Introduction

This document presents a mathematical breakdown of uncertainty in regression modeling. We analyze uncertainty in:

- **Input data** (*epistemic uncertainty*) – unknown measurement errors.
- **Model structure** (*epistemic uncertainty*) – choice of linear regression.
- **Parameter estimation** (*epistemic uncertainty*) – coefficients are estimated from finite samples.
- **Target data (response variability)** (*aleatory uncertainty*) – inherent randomness in outcomes.
- **Intrinsic randomness** (*aleatory uncertainty*) – stochastic noise added to the system.

2 Synthetic Data Generation

We generate synthetic data for a multivariate regression model with **known** uncertainty in input data, model structure, parameters, target data, and intrinsic randomness. We assume the true relationship follows a linear model with log-normal distributions for strictly positive data:

$$X_i \sim \text{LogNormal}(\mu_X, \sigma_X^2), \quad i = 1, \ldots, N \tag{1}$$

$$\beta_0, \beta_1 \sim \text{Normal}(\mu_\beta, \sigma_\beta^2) \tag{2}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2) \quad \text{(aleatory uncertainty)} \tag{3}$$

The true response is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4}$$

where ϵ_i represents the irreducible noise in the system (aleatory uncertainty).

2.1 Generating Predictor Variables

We define two independent predictor variables X_1 and X_2, sampled from log-normal distributions:

$$X_1 \sim \text{LogNormal}(\mu_1, \sigma_1) \tag{5}$$

$$X_2 \sim \text{LogNormal}(\mu_2, \sigma_2) \tag{6}$$

where $\mu_1 = 1$, $\sigma_1 = 0.2$, and $\mu_2 = 1.5$, $\sigma_2 = 0.3$.

2.2 True Model with Known Parameters

The true response variable Y follows:

$$Y = \exp\left(\beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2\right) \cdot \epsilon \tag{7}$$

where:

- $\beta_0 \sim \mathcal{N}(2, 0.5)$ is the intercept,
- $\beta_1 \sim \mathcal{N}(1.2, 0.2)$ and $\beta_2 \sim \mathcal{N}(-0.8, 0.15)$ are regression coefficients,
- $\epsilon \sim \text{LogNormal}(0, 0.1)$ is the multiplicative noise.

3 Deterministic Model

We fit a **deterministic** linear regression model to the log-transformed data:

$$\log Y = \alpha_0 + \alpha_1 \log X_1 + \alpha_2 \log X_2 + \epsilon \tag{8}$$

where α_i are estimated via **ordinary least squares (OLS)**:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{9}$$

where $\mathbf{X}$ is the design matrix and $\mathbf{y}$ is the vector of observed values $\log Y$.

3.1 Deterministic Predictions

$$\hat{Y}_{\text{det}} = \exp(\hat{\alpha}_0 + \hat{\alpha}_1 \log X_1 + \hat{\alpha}_2 \log X_2) \tag{10}$$

3.2 Coefficient of Determination ($\mathbf{R^2}$)

The goodness of fit for the deterministic model is given by:

$$R^2_{\text{det}} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_{\text{det},i})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{11}$$

4 Stochastic Ensemble Model

To incorporate uncertainty, we generate **multiple simulations** of the model using Monte Carlo sampling.

4.1 Monte Carlo Sampling

For each realization $k = 1, \dots, L$:

$$X_1^{(k)} \sim \text{LogNormal}(\log X_1, 0.1) \tag{12}$$

$$X_2^{(k)} \sim \text{LogNormal}(\log X_2, 0.15) \tag{13}$$

The regression parameters are also sampled:

$$\beta_0^{(k)} \sim \mathcal{N}(\hat{\alpha}_0, 0.3) \tag{14}$$

$$\beta_1^{(k)} \sim \mathcal{N}(\hat{\alpha}_1, 0.1) \tag{15}$$

$$\beta_2^{(k)} \sim \mathcal{N}(\hat{\alpha}_2, 0.1) \tag{16}$$

Each realization of Y is computed as:

$$Y_{\text{sim}}^{(k)} = \exp(\beta_0^{(k)} + \beta_1^{(k)} \log X_1^{(k)} + \beta_2^{(k)} \log X_2^{(k)}) \tag{17}$$

with additional stochastic noise:

$$Y_{\text{sim}}^{(k)} = Y_{\text{sim}}^{(k)} \cdot \text{LogNormal}(0, 0.1) \tag{18}$$

4.2 Ensemble Mean Prediction

The final stochastic model prediction is computed as the **mean across all simulations**:

$$\hat{Y}_{\text{stoch}} = \frac{1}{L} \sum_{k=1}^{L} Y_{\text{sim}}^{(k)} \tag{19}$$

4.3 Stochastic Model R²

The coefficient of determination for the ensemble model is computed as:

$$R_{\text{stoch}}^2 = \left[\frac{\text{Cov}(\log Y, \log \hat{Y}_{\text{stoch}})}{\sigma_Y \sigma_{\hat{Y}_{\text{stoch}}}} \right]^2 \tag{20}$$

5 Visualization and Comparison

We generate three plots to compare results:

5.1 Plot 1: Deterministic Model

A scatter plot of observed vs. deterministic predictions:

$$\log Y_{\text{true}} \text{ vs. } \log \hat{Y}_{\text{det}}$$

with a regression line.

5.2 Plot 2: Stochastic Ensemble Model

A scatter plot of observed vs. ensemble predictions:

$$\log Y_{\text{true}} \text{ vs. } \log \hat{Y}_{\text{stoch}}$$

with uncertainty bands (confidence intervals from simulations).

5.3 Plot 3: Comparison of Deterministic vs. Stochastic

Both deterministic and stochastic predictions are plotted together:

- **Blue Line:** Deterministic best-fit regression
- **Red Line:** Stochastic ensemble mean regression

R^2 values are displayed for both models.

6 Conclusion

- The **deterministic model** provides a single best-fit prediction but does not account for **uncertainty** in inputs, parameters, or observations.
- The **stochastic model** provides a distribution of possible predictions, yielding a more robust estimate of the true system.
- The **ensemble mean prediction** can be more accurate when uncertainty is properly modeled.
- **Key Insight:** Stochastic modeling provides a **realistic representation of uncertainty** compared to a single deterministic best-fit.