1 Glossary

- Explanatory variable (input variable, predictor, covariate): Often written as x. Generally written as x if the data has been observed, $\tilde{x}$ if it hasn't.
- Outcome variable (output variable, target, response): Often written as y. Generally written as y if the data has been observed, $\tilde{y}$ if it hasn't.
- Data: The combination of input and output variables that have been observed. $D = (x_i, y_i)$.
- Model: Some way (parametric or non) of relating the input and output variables. $y = M(x, \theta)$.
- Likelihood: The liklihood of the data given the model (M), its parameters (θ); $\mathcal{L}(D|M, \theta)$
- Prior probability distribution: The probability of the parameters (θ) for the model, before conditioning on the data. This maybe be determined by previous work, or natural limit (some properties, such as variance, must be > 0).
- Posterior probability distribution: The probability of the parameters (θ) for the model, given the data. $p(\theta|D, M)$
- Prior predictive distribution: The prediction if y for x, given a model and a prior distribution over the parameters for that model. $p(y|x) = \int p(y|x, \theta)p(\theta)d\theta$.
- Posterior predictive distribution: The prediction of $\tilde{y}$ for $\tilde{x}$, given a model conditioned on some data; $p(\tilde{y}|\tilde{x}, D, M)$. In the process of conditioning on the data, we obtain a posterior on the model parameters. Thus, $p(\tilde{y}|\tilde{x}, D, M) = \int p(\tilde{y}|\tilde{x}, \theta, M)p(\theta|D, M)d\theta$, or the probaility of $\tilde{y}$ at a given θ, weighted by the posterior probability of that θ.
- Latent variable

Let's give a full example. Consider subhalo abundance matching, a method of populating dark matter halos with stellar mass. Our target, y, is the stellar mass. We choose an explanatory variable (such as halo mass) that we thing correlates well (covaries) with this - this is x. We have some model M to compute $y|x$. This model parameterizes the SMHM relation and its scatter $y = M(x, \theta)$. To find those other parameters, we find the liklihood of our data given the model and θ. To do this, we also assume a cosmology and that an N-body simulation is a good realization of the universe. The Likelihood compares the summary statistic of the N-body + model and observed data.