**Linear Vs. Bayesian Regression**

**Linear Regression**

In general, frequentists think about Linear Regression as follows:

Y = Xβ + ϵ

where Y is the output we want to predict (or *dependent* variable), X is our predictor (or *independent* variable), and β are the coefficients (or parameters) of the model we want to estimate. ϵ is an error term which is assumed to be normally distributed.

We can then use Ordinary Least Squares or Maximum Likelihood to find the best fitting β.

**Probabilistic Reformulation**

Bayesians take a probabilistic view of the world and express this model in terms of probability distributions. The above linear regression can be rewritten to yield:

Y ∼ N (Xβ, σ2)

In words, we view Y as a random variable (or random vector) of which each element (data point) is distributed according to a Normal distribution. The mean of this normal distribution is provided by the linear predictor with variance σ2.

While this is essentially the same model, there are two critical advantages of Bayesian estimation:

* Priors: We can quantify any prior knowledge we might have by placing priors on the parameters. For example, if we think that σ is likely to be small we would choose a prior with more probability mass on low values.
* Quantifying uncertainty: We do not get a single estimate of β as Linear Regression but instead a complete posterior distribution about how likely different values of β are. For example, with few data points our uncertainty in β will be very high and we'd be getting very wide posteriors.

**Hierarchical Regression Model**

**Possibilities we can think of:**

**Pooling of measurements**

That's the easy option considering parameters don’t vary for different treatments. We'll just pool all the data and estimate one big regression to assess the influence of parameters for all treatments together.

Critically, we are only estimating *one* parameter set (k1, Y1, af1, as1, sf1) for all measurements over all treatments.

**Separate regressions**

But what if we are interested whether different treatments actually have different parameters? Then we could just estimate n (number of treatments) different parameter set -- one for each treatment.

This is the extreme opposite model, where above we assumed all treatments are exactly the same, here we are saying that they share no similarities whatsoever which ultimately is also questionable.

**Hierarchical Regression: The best of both worlds**

Fortunately there is a middle ground to both of these extreme views. Specifically, we may assume that while parameters are different for each treatment, but they all come from a common group distribution:

kt∼(μk, σk2); Yt∼(μY, σY2); so on…

We thus assume the parameters to come from a normal distribution centered around their respective group mean μ with a certain standard deviation σ2, the values (or rather posteriors) of which we also estimate. That's why this is called multilevel or hierarchical modeling.

Further reading:

<http://twiecki.github.io/blog/2014/03/17/bayesian-glms-3/#The-best-of-both-worlds:-Hierarchical-Linear-Regression-in-PyMC3>

A nice paper for explanation on hierarchical linear modeling:

<http://www.tqmp.org/RegularArticles/vol08-1/p052/p052.pdf>