

Posterior Distribution

• The output of a Bayesian regression is not a set of best fit parameters, but a probability distribution for each parameter and σ<sub>ε</sub> (called the posterior distribution)

• How de we interpret the posterior distribution?

• We need to summarize the distribution:

– Use the mode, mean, median, or range midpoint as an equivalent "best estimate" of the parameter

– Use the distribution to calculate "credible interval" (quantiles) for the parameter

Posterior Distribution

• Posterior distribution describes how much the data has changed our prior beliefs

• Bernstein-von Mises Theorem: for a sufficiently large sample size, the posterior distribution becomes independent of the prior distribution (so long as the prior is not either 0 or 1)

- The posterior tends towards a normal distribution with a mean equal to the MLE (assuming iid data), a restatement of the central limit theorem

- The effect of the prior diminishes as the amount of data increases

Prior Distribution

• The prior distribution  $\mathbb{P}(\theta)$  is really shorthand notation for  $\mathbb{P}(\theta|I)$ , where I is all the information we have about the problem before we start collecting data

• If we have NO information about what the parameters could or should be, then  $\mathbb{P}(\theta|I)$  is a constant (called an uninformative prior or objective prior), and the posterior distribution equals the likelihood function

• We almost always have some information

Prior Choice

• Non-informative (baseline or objective) prior

- Ex: a uniform probability over the expected range of possible values

- Flat priors are not always uninformative! Ex: should we have a uniform distribution of slopes, or uniform distribution of the angle of the line, or its sine?

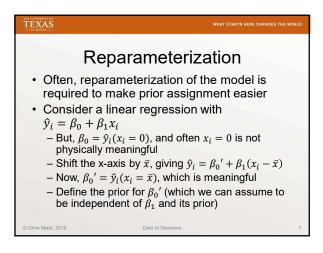
• Substantive (informative) prior

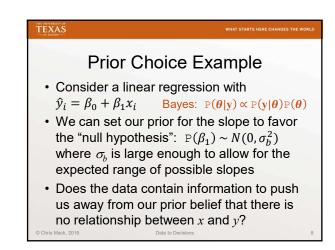
- Use some problem-specific information to provide a prior distribution for each model parameter

- Based on previous data, experiments, knowledge

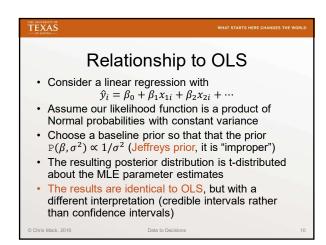
- Sometimes one can assume the prior for each parameter is independent of the others (P(θ) = P(β)P(σ²)), but frequently a joint probability distribution is required

- Setting the prior to a delta function fixes a parameter independent of the data (we never do this in general)





Conjugate Priors
 For a given likelihood distribution, analytical solutions of the posterior distribution are only possible for special cases of priors (called conjugate priors)
 Example: For iid normal errors, the conjugate prior for β is normal, and for σ² is inverse gamma
 Usually, we need to solve Bayes' equation numerically
 Markov Chain Monte Carlo (MCMC) sampling
 Result is a set of points from the posterior distribution that we then summarize (mean, or maximum a posteriori – MAP – estimate of the mode)



Interpretation

• 95% Confidence Interval (frequentist)

- Our parameter is an unknown constant, and with a large number of repeated samples, 95% of such calculated confidence intervals would include the true value of the parameter

• 95% Credible Interval/Region (Bayesian)

- Our parameter is a random variable, with a 95% probability of falling within the given interval

Lecture 75: What have we learned?

How is the posterior distribution used (summarized) to tell us about model parameters?

What does the Bernstein—von Mises theorem tell us about the relationship between the prior and posterior distributions?

What is the difference between uninformative and substantive priors?

What is the Jeffreys prior and how does it apply to linear regression?

Explain the difference between confidence intervals and credible intervals