CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 75
Bayesian Regression, part 2

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Bayesian Regression

$$\underbrace{P(\theta|\mathbf{y})}_{\text{Posterior Distribution}} = \frac{\overbrace{P(\mathbf{y}|\theta)}^{\text{Likelihood Function}}\overbrace{P(\theta)}^{\text{Prior Distribution}}}{\underbrace{P(\mathbf{y})}_{\text{Normalizing constant}}}$$

- We already know how to calculate the likelihood function (we make assumptions about the pdf of $\mathbf{y}$)
- The term $P(\mathbf{y})$ is a constant (independent of θ) and used to normalize the posterior pdf
- Prior distribution $P(\theta)$: our understanding of the model parameters and σ_ε before we began the experiment
- How do we interpret the posterior distribution?

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_ε (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 $P(\theta)$ is really shorthand notation for $P(\theta|\mathcal{I})$, where $\mathcal{I}$ 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 $P(\theta|\mathcal{I})$ 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 ($P(\boldsymbol{\theta}) = P(\boldsymbol{\beta})P(\sigma^2)$), 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)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Reparameterization

- Often, reparameterization of the model is required to make prior assignment easier
- Consider a linear regression with $\hat{y}_i = \beta_0 + \beta_1 x_i$
  - But, $\beta_0 = \hat{y}_i(x_i = 0)$, and often $x_i = 0$ is not physically meaningful
  - Shift the x-axis by $\bar{x}$, giving $\hat{y}_i = \beta_0' + \beta_1(x_i - \bar{x})$
  - Now, $\beta_0' = \hat{y}_i(x_i = \bar{x})$, which is meaningful
  - Define the prior for β_0' (which we can assume to be independent of β_1 and its prior)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Prior Choice Example

- Consider a linear regression with $\hat{y}_i = \beta_0 + \beta_1 x_i$ Bayes: $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y}) \propto \mathrm{P}(\mathbf{y}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})$
- We can set our prior for the slope to favor the "null hypothesis": $\mathrm{P}(\beta_1) \sim N(0, \sigma_b^2)$ where σ_b is large enough to allow for the expected range of possible slopes
- Does the data contain information to push us away from our prior belief that there is no relationship between x and y?

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

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 σ^2 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)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Relationship to OLS

- Consider a linear regression with
$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$$
- Assume our likelihood function is a product of Normal probabilities with constant variance
- Choose a baseline prior so that that the prior $\mathrm{P}(\beta, \sigma^2) \propto 1/\sigma^2$ (Jeffreys prior, it is "improper")
- The resulting posterior distribution is t-distributed about the MLE parameter estimates
- The results are identical to OLS, but with a different interpretation (credible intervals rather than confidence intervals)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

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

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

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