Bayesian Methods for Regression in R

Nels Johnson

Lead Collaborator, Laboratory for Interdisciplinary Statistical Analysis

03/14/2011
Outline

- What is Linear Regression?
- Intro to Bayesian Statistics
- More on Priors
- Bayesian Estimation and Inference
- Computing Issues
- Examples
Suppose you have a variable $Y$ whose outcome is considered random.

Examples: Amount of soil erosion at a particular site; How much time it takes for a fungus to develop a resistance to a fungicide; How many feet a moving car will need to come to a complete stop.

Suppose you have some other other variables $X$ which are treated as fixed.

Examples: Kind of plants are growing in the soil; The composition of the fungicide; How fast the car is traveling.

If you think the observed value of $Y$ is dependent on $X$ then regression is a tool for modeling that dependence.
Linear Regression

- Linear regression assumes that the underlying structure of $Y$ is as linear combination of the variables $X$.
- Example:
  \[ Y_{dist} = \beta_0 + X_{\text{speed}} \beta_{\text{speed}} + \epsilon \]
- Often the formula for linear regression will be condensed using matrix algebra:
  \[ Y = X\beta + \epsilon \]
- Since $Y$ is considered a random variable, we will naturally expect it not to follow the underlying structure exactly.
- $\epsilon$ is the term of the model that describes $Y$’s random variation around it’s underlying structure $X\beta$.
- The most common way to describe the random variation is with $\epsilon \sim N(0, \sigma^2)$. 
Estimating Parameters

- Since we don’t know the values for $\beta$ and $\sigma^2$ we’ll need to estimate them based on our data ($Y$, $X$).
- Traditionally we would do so by finding the $\beta$ and $\sigma^2$ that maximize the likelihood:

$$L(Y_i | X_i, \beta, \sigma^2) = \prod_{i=1}^{n} N(Y_i | X_i \beta, \sigma^2)$$

- Note: the likelihood is also the joint distribution of the data.
- The estimates $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $s^2 = \frac{(Y - X \hat{\beta})^T (Y - X \hat{\beta})}{n-p}$ would be considered random variables because they are functions of the random variable $Y_i$. 
Interpretation of Parameters

- We might compute a confidence interval for $\beta$ or perform a test to see if it is significantly different from zero.
- Confidence intervals are interpreted in terms of the proportion (i.e. relative frequency) of times they should capture the true parameter in the long run.
- This is because it is the endpoints of the interval that are random variables and the parameter is fixed. The interval either captures the fixed parameter or it doesn’t.
- Because of this interpretation, this paradigm of statistics is called the frequentist paradigm (or classical paradigm).
Example 1.2

- This is an example of regression using the `lm()` function in R.
- Go to Example 1.2 in “Bayes Reg 2011.r”
The Bayesian paradigm is named after Rev Thomas Bayes for its use of his theorem.

Take the rule for conditional probability for two events $A$ and $B$:

$$ P(A|B) = \frac{P(A \cap B)}{P(B)} $$

Bayes discovered that this is equivalent to:

$$ P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\int P(B|A)P(A)dA} $$

This is known as Bayes’ Theorem or Bayes’ Rule.
The mathematician Pierre-Simon Laplace got the idea that instead of just defining probability on variables, we could also define probability on parameters too. And by using Bayes’ Rule we can make inference on parameters. Effectively treating parameters as random variables.

In our regression example, let $\theta = \{\beta, \sigma^2\}$ and $D =$ the data. Using Bayes’ Rule we get:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(\theta|D)$ is called the **posterior distribution**. It is what we will use to make inference about the parameters $\theta = \{\beta, \sigma^2\}$. 
The mathematician Pierre-Simon Laplace got the idea that instead of just defining probability on variables, we could also define probability on parameters too. And by using Bayes’ Rule we can make inference on parameters. Effectively treating parameters as random variables.

In our regression example, let $\theta = \{\beta, \sigma^2\}$ and $D =$ the data. Using Bayes’ Rule we get:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(D|\theta)$ is the likelihood we discussed previously. It contains all the information about $\theta$ we can learn from the data.
Bayesian Paradigm

The mathematician Pierre-Simon Laplace got the idea that instead of just defining probability on variables, we could also define probability on parameters too. And by using Bayes’ Rule we can make inference on parameters. Effectively treating parameters as random variables.

In our regression example, let $\theta = \{\beta, \sigma^2\}$ and $D =$ the data. Using Bayes’ Rule we get:

$$P(\theta | D) = \frac{P(D | \theta) P(\theta)}{P(D)}$$

$P(\theta)$ is called the prior distribution for $\theta$. It contains the information we know about $\theta$ before we observe the data.
The mathematician Pierre-Simon Laplace got the idea that instead of just defining probability on variables, we could also define probability on parameters too. And by using Bayes’ Rule we can make inference on parameters. Effectively treating parameters as random variables.

In our regression example, let \( \theta = \{ \beta, \sigma^2 \} \) and \( D = \) the data. Using Bayes’ Rule we get:

\[
P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}
\]

\( P(D) \) is the normalizing constant of the function \( P(D|\theta)P(\theta) \) such that \( P(\theta|D) \) is a proper probability distribution.
• Often knowing the normalizing constant $P(D)$ is not necessary for Bayesian inference.

• When it’s removed we get:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \Leftrightarrow \text{posterior} \propto \text{likelihood} \times \text{prior}$$

• The $\propto$ symbol means that $P(\theta|D)$ is proportional to (i.e. a scalar multiple of) $P(D|\theta)P(\theta)$.

• In this case that scalar is $P(D)$.

• Usually $P(D|\theta)$ and $P(\theta)$ only need to be known up to proportionality as well because their normalizing constants get lumped in with $P(D)$.
How It Works:

- The advantage of Bayesian analysis is sequentially updating beliefs about $\theta$.
- Process: Prior belief about $\theta \rightarrow$ Observe data $\rightarrow$ update belief about $\theta$.
- That is $P(\theta) \rightarrow P(D|\theta)P(\theta) \rightarrow P(\theta|D)$.
- Now that we’ve done one experiment and have $P(\theta|D_1)$ we can conduct another experiment using $P(\theta|D_1)$ as the prior for $\theta$.
- Leads us to:

$$P(\theta|D_2, D_1) \propto P(D_2|\theta)P(\theta|D_1)$$

- This process can be continually repeated.
- Example: Last year you did a study relating levels of heavy metals in streams to the size of mussels in the stream. Use the posterior from last year’s study as your prior for this year’s study.
Prior Distributions

- Just like how you need to specify the likelihood for the problem, you need to specify the prior distribution.
- This is not an intuitive process for many people at first mainly because of the prior.
- What helped me the most: think of the prior as a measure of our uncertainty in the $\theta$'s true value.
- Very often it is easiest to write the joint prior as independent univariate priors, e.g. $P(\theta) = P(\beta)P(\sigma^2)$.
- Some more important terminology that comes up when talking about priors:
  - Informative and noninformative priors
  - Proper and improper priors
  - Conjugate priors
  - Reference priors
If we use maximum likelihood to estimate $\beta$, with $\sigma^2$ known, then $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. We also know the sampling distribution for $\hat{\beta}$ is normal.

It turns out, if we can pick the prior $\pi(\beta) \propto 1$, then $\pi(\beta | X, Y) \sim N((X^T X)^{-1} X^T Y, \sigma^2 (X^T X)^{-1})$.

This is an example improper prior, since it isn't a valid probability distribution, but the posterior $\pi(\beta | X, Y)$ is.

It is also considered an uninformative.

In fact, it is so uninformative it's called the reference prior, since it's the least informative prior we could pick (under some criterion I don't want to get into).
Example 1.3

- This example will allow us to see how the prior is updated after we see data.
- Key points:
  - As the sample size increases, impact of prior on posterior decreases.
  - As the variance of the prior increases (i.e. diffuse prior), impact of the prior on the posterior decreases.
  - Where the prior is centered is not important if the prior is diffuse enough.
If $P(\theta)$ is a conjugate prior for $\theta$ with likelihood $P(D|\theta)$, then $P(\theta|D)$ has the same family of distributions as the prior.

This is nice because it means that in some cases we can pick a prior such that the posterior distribution will be a distribution we understand well (e.g. normal, gamma).

This means computational methods might not be needed for inference or that the computation methods will be fast and simple to implement.

For a list of some: http://en.wikipedia.org/wiki/Conjugate_prior
Common Conjugate Examples

\[ Y = X\beta + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2 I) \text{ or } \epsilon \sim N(0, \Sigma) \]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Likelihood</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>Inverse-Gamma</td>
<td>Normal</td>
<td>Inverse-Gamma</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>Inverse-Wishart</td>
<td>Normal</td>
<td>Inverse-Wishart</td>
</tr>
<tr>
<td>( \sigma^{-2} )</td>
<td>Gamma</td>
<td>Normal</td>
<td>Gamma</td>
</tr>
<tr>
<td>( \Sigma^{-1} )</td>
<td>Wishart</td>
<td>Normal</td>
<td>Wishart</td>
</tr>
</tbody>
</table>
The rest of this presentation will be in R.
I’ve left all my slides from an older version of this talk for those that may be interested in reading them later.
Bayesian inference is usually done on the joint posterior distribution of all the parameters $P(\theta|D)$.

Sometimes it is done on the marginal distribution of a single parameter. Such as:

$$P(\beta|D) = \int P(\theta|D) d\sigma^2$$

Because the posterior is a probability distribution for $\theta$, performing inference on $\theta$ is as simple as finding relevant summaries based on its posterior.
Two very popular point estimators for $\theta$:

- The mean of the posterior distribution (posterior mean).
- The maximum a posteriori estimator also known as the MAP estimator.

The MAP estimator is $\arg\max_{\theta} P(\theta|D)$.

You could use any measure of center that makes sense for $P(\theta|D)$. 
Bayesian confidence intervals are called credible intervals. So a 95% Bayesian confidence interval is called a 95% credible interval.

There are two popular ways to find them:
- Highest posterior density, often shortened to HPD.
- Equal tail probability.

A 95% HPD credible interval is the smallest interval with probability 0.95.

A 95% Equal tail probability interval use the values that have cumulative probability of 0.025 and 0.975 as the endpoints. This is not preferred when the posterior is highly skewed or multimodal.
Hypothesis Testing

- Since we can talk about the probability of $\theta$ being in some interval, this makes interpretation of certain kinds of hypothesis tests much easier.

- For instance $H_0 : \beta \leq 0$ vs. $H_a : \beta \geq 0$ You can find $P_{\beta|D}(\beta \leq 0)$ and if it is sufficiently small then reject the null hypothesis in favor of the alternative.

- We could also test $H_0 : a \leq \beta \leq b$ vs $H_0 : \beta \leq a$ or $b \leq \beta$ where $a$ and $b$ are chosen such that if $H_0$ were true then $\beta$ would have no practical effect. Again, compute $P_{\beta|D}(H_0)$ and if it is sufficiently small, reject $H_0$.

- Some special steps need to be take when trying to test $H_0 : \beta = 0$ which are beyond the scope of this course.

- Also instead of rejecting $H_0$ or accepting $H_a$, you could just report the evidence (i.e. the probability) for the reader to decide.
Sampling From the Joint Posterior

- As stated previously we want to perform inference on $\theta$ using the joint posterior $P(\theta|D)$.
- Summarizing the joint posterior may be difficult (or impossible).
- A solution to this problem is to use a computer algorithm to sample observations from the joint posterior.
- If we have a large enough sample from $P(\theta|D)$ then the distribution of the sample should approximate $P(\theta|D)$.
- This means we can just summarize the sample from $P(\theta|D)$ to make inference on $\theta$ instead of using the distribution $P(\theta|D)$ directly.
The big three algorithms are as follows:
- Gibbs sampling.
- Metropolis algorithm.
- Metropolis-Hastings algorithm.

All three of these algorithms are Markov Chain Monte Carlo algorithms. Commonly known as MCMC algorithms.

What that means for us is that after you run the algorithm “long enough” you will eventually start sampling from $P(\theta|D)$.

The Metropolis-Hasting algorithm is a generalization of the other two and it so similar to the Metropolis algorithm that the two titles are often interchanged.
The Gibbs sampler is a much simplified version of the Metropolis-Hastings algorithm. So simplified in fact it isn’t obvious at all that they are related.

All three algorithms require you to know the conditional posterior distribution of each parameter in the model (e.g. $P(\beta|D, \sigma^2)$ and $P(\sigma^2|D, \beta)$).

The Gibbs sampler requires that the user can sample from each of the conditional posterior distributions using computer software.

This is where conjugate priors become very popular. If the conditional posterior distributions are all normal, gamma, or some other distribution we can readily sample from, then the Gibbs algorithm makes sampling from the joint posterior very easy.

Other details about this algorithm are beyond the scope of this course.

The data analysis examples we will discuss both use a Gibbs sampler.