Bayesian Methods for Regression in R

Nels Johnson

Lead Collaborator, Laboratory for Interdisciplinary Statistical Analysis
Department of Statistics, Virginia Tech

03/13/2012

When would I want to use a regression method?

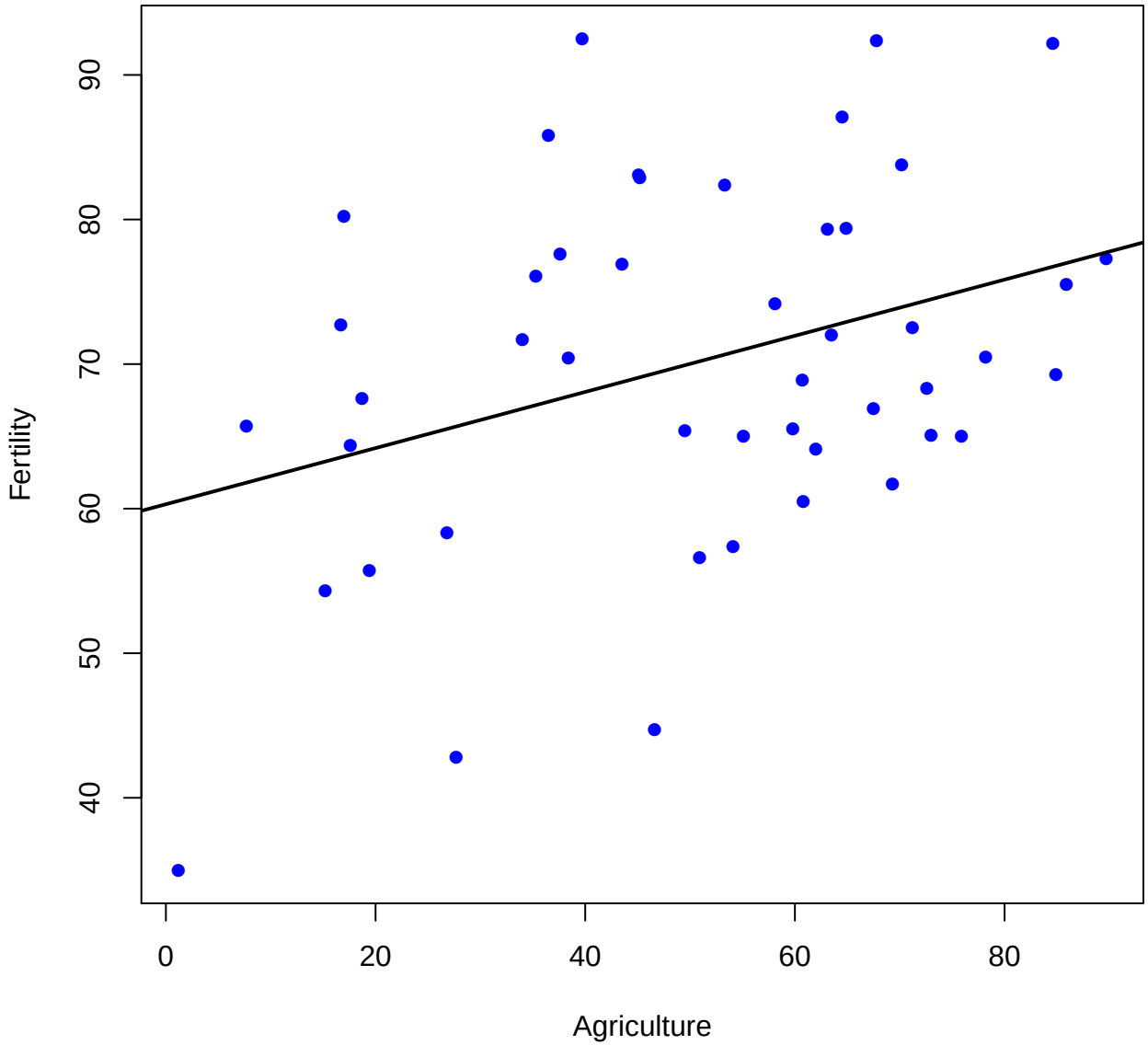
- Regression methods try to explain the relationship between two sets of variables Y and X.
- Y is considered random and X is considered fixed.
- Examples:
  - Fertility rates in Switzerland based on economic factors.
  - Number of insects killed based on insecticide.

Linear Regression

- A common way to describe the relationship between Y and X is to say Y is a linear combination of X plus some error:

$$\begin{aligned} Y &= \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \epsilon \\ &= X\beta + \epsilon \end{aligned}$$

- Often it is reasonable to assume $\epsilon \sim N(0, \sigma^2)$.
- Given our data Y and X, our goal is then to determine reasonable values for β and σ^2, and then describe our uncertainty in them.


Fertility
Agriculture
90
80
70
60
50
40
0
20
40
60
80

Maximum Likelihood Estimation

- Traditionally, we find reasonable values for β and σ^2 using maximum likelihood estimation (MLE).
- The likelihood, $L(\beta, \sigma^2|Y, X)$, is the same as the joint distribution of Y, $P(Y|X, \beta, \sigma^2)$.
- When we assume $\epsilon \sim N(0, \sigma^2)$, then $Y \sim N(X\beta, \sigma^2)$.
  - $\hat{\beta}_{\text{MLE}} = (X^T X)^{-1} X^T Y$
  - $\widehat{\text{var}(\hat{\beta}_{\text{MLE}})} = s^2 (X^T X)^{-1}$
  - $s^2 = (Y - X\hat{\beta}_{\text{MLE}})^T (Y - X\hat{\beta}_{\text{MLE}})/(n - p) = \text{MSE}$

R examples sections 0 and 1

Outline: Lecture

- What is Bayes' Rule?
  - What is the likelihood?
  - What is the prior distribution?
    - How should I choose it?
    - Why use a conjugate prior?
    - What is an subjective versus objective prior?
  - What is the posterior distribution?
    - How do I use it to make statistical inference?
    - How is this inference different from frequentist/classical inference?
    - What computational tools do I need in order to make inference?

Why use Bayes?

- Ease of interpretation. Interpretation of the posterior probability as a measure of evidence.
  - Credible interval vs confidence interval.
  - Posterior probability vs p-values.
- Does not rely on assumption that sample is large.
  - Generalized linear models
  - Mixed models
  - Nonlinear models
- It lends itself well to the sequential nature of experimentation.
  - Prior knowledge + Data $\rightarrow$ Posterior knowledge
  - Old posterior knowledge + New data $\rightarrow$ New posterior knowledge

Bayes' Theorem

- 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.

A little history

- The mathematician Pierre-Simon Laplace popularized 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. He laid the groundwork for the Bayesian paradigm of statistics.

Bayesian Paradigm

- 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 $\theta = \{\beta, \sigma^2\}$.

Bayesian Paradigm

- 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 $\theta = \{\beta, \sigma^2\}$.
- $P(D|\theta)$ is the likelihood we discussed previously. It contains all the information about θ we can learn from the data.

Bayesian Paradigm

- 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 $\theta = \{\beta, \sigma^2\}$.
- $P(D|\theta)$ is the likelihood we discussed previously. It contains all the information about θ we can learn from the data.
- $P(\theta)$ is called the prior distribution for θ. It contains the information we know about θ before we observe the data.

Bayesian Paradigm

- 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 $\theta = \{\beta, \sigma^2\}$.
- $P(D|\theta)$ is the likelihood we discussed previously. It contains all the information about θ we can learn from the data.
- $P(\theta)$ is called the prior distribution for θ. It contains the information we know about θ before we observe the data.
- $P(D)$ is the normalizing constant of the function $P(D|\theta)P(\theta)$ such that $P(\theta|D)$ is a proper probability distribution.

Bayesian Inference

- The posterior distribution $P(\theta|D)$ is what we use to make inference on the parameters given the data.
- Any way you might normally summarize a distribution, you can use to summarize a parameter.
- Density plots of $P(\theta|D)$ are a great place to start.
- For point estimates:
  - Posterior mean: $E_\theta[P(\theta|D)]$.
  - Posterior median: $P_{0.5}(\theta|D)$.
  - Maximum *a posteriori* (MAP): $\arg\max_\theta[P(\theta|D)]$.
- $(1-\alpha)100\%$ credible intervals (Bayesian confidence intervals):
  - Equal tail interval: $[P_{\alpha/2}(\theta|D), P_{1-\alpha/2}(\theta|D)]$.
  - Highest Posterior Density (HPD): Smallest interval to cover $(1-\alpha)100\%$ of $P(\theta|D)$.
  - We'll talk more about these in the software section.

Bayesian hypothesis testing

- Interval-based hypothesis tests are very easy to make:
  - $H_0 : \beta \leq 0$ vs $H_1 : 0 < \beta$
  - $H_0 : -0.1 \leq \beta \leq 0.1$ vs $H_1 : \beta < -0.1$ or $0.1 < \beta$
- Simply find $P(H_0)$ or $P(H_1)$
  - $P(\beta \leq 0 | X, Y)$
  - $P(-0.1 \leq \beta \leq 0.1 | X, Y)$

Bayesian hypothesis testing

- Others are much harder:
  - Suppose $\beta = \{\beta_1, \beta_2\}$
  - $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- This is actually a model selection problem.
- Topic too advanced for here, but three ways to do model selection as a Bayesian:
  - Posterior model probability.
  - Bayes factors.
  - Select prior designed for model selection.

Selecting a prior distribution

- For illustrative purposes, let's pretend we know σ^2 and we want to choose a prior for β.
- Two Examples (and popular choices):
  - $P(\beta) \propto 1$
  - $P(\beta) \sim N(0, t^2)$
- What is the posterior $P(\beta|Y, X, \sigma^2)$?

Example 1

- For $P(\beta) \propto 1$:

$$\begin{aligned} P(\beta|Y, X, \sigma^2) &= N(\text{mean}, \text{var}) \\ \text{mean} &= (X^T X)^{-1} X^T Y \\ \text{var} &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

- If you recall from earlier, this is the same as the MLE, when σ^2 is known.
- $P(\beta) \propto 1$ is not a proper probability distribution, but it is OK since $P(\beta|-)$ is a proper probability distribution.
- When $P(\theta)$ is not a proper probability distribution it is called an improper prior.
- Care should be taken when selecting improper priors, since they won't always lead to proper posteriors.

Example 2

- For $P(\beta) \sim N(0, t^2 I)$:

$$\begin{aligned} P(\beta|Y, X, \sigma^2) &= N(\text{mean}, \text{var}) \\ \text{mean} &= [\sigma^{-2} X^T X + t^{-2} I]^{-1} X^T Y \sigma^{-2} \\ \text{var} &= [\sigma^{-2} X^T X + t^{-2} I]^{-1} \end{aligned}$$

- This is the same answer we'd get if we used the MLE for ridge regression.
- Ridge regression is a penalized regression method that places a penalty of the L_2-norm of β.
- The estimates are biased, but have smaller variance.
- When the resulting posterior is the same distribution as the prior, the prior is called as a conjugate prior.

Comments

- In general, the choice of prior on θ can be thought of as choosing a penalty on θ.
- The strength of a penalty is relative to the sample size N.
- The larger the sample size, the more information will come from the likelihood than from the prior.
- The stronger the penalty, the stronger our "prior belief".
- Strong priors sometimes called informative or subjective, and weak prior are called uninformative or objective.
- Note: All priors contain information. There is no such thing as a truly "uninformative" prior.

R examples section 2

What about the variance?

- Let's now assume that both β and σ^2 are unknown.
- Two more examples:
  - $P(\beta, \sigma^2) = P(\beta)P(\sigma^2) \propto \sigma^{-2}$
  - $P(\beta, \sigma^2) = P(\beta)P(\sigma^2) \sim N(\beta; 0, t^2)IG(\sigma^2; a, b)$
- IG is shorthand for Inverse-Gamma distribution.
- If $X \sim \text{Gamma}(\text{shape} = a, \text{rate} = b)$, then $Y = 1/X \sim IG(a, b)$

Example 3

- For $P(\beta, \sigma^2) = P(\beta)P(\sigma^2) \propto \sigma^{-2}$, let's do something a little different.
- Instead of doing inference on $P(\beta, \sigma^2|X, Y)$, let's do it on the marginal distribution of β, $P(\beta|X, Y)$.

$$\begin{aligned} P(\beta|Y, X) &= \int P(\beta|\sigma^2, Y, X)d\sigma^2 \\ &= \int N[(X^TX)^{-1}X^TY, \sigma^2(X^TX)^{-1}]d\sigma^2 \\ &= T_{df=N-p}(\text{mean}, \text{var}) \\ \text{mean} &= (X^TX)^{-1}X^TY \\ \text{var} &= s^2(X^TX)^{-1} \end{aligned}$$

- This is analogous the the distribution of $\hat{\beta}_{\text{MLE}}$ discuss earlier.

Example 4

- For $P(\beta, \sigma^2) = P(\beta)P(\sigma^2) \sim N(\beta; 0, t^2)IG(\sigma^2; a, b)$, $P(\beta, \sigma^2|Y, X)$ starts to get more complicated.
- It is actually a distribution called the Normal-Inverse-Gamma distribution. But let's pretend we didn't know that.
- How would we perform inference on a posterior distribution when we do not know it's functional form?
- This means we don't know its mean, variance, quantiles etc.
- This problem is ubiquitous in Bayesian inference.
- Solution: Markov chain Monte Carlo (MCMC) methods.

Markov chain Monte Carlo

- It turns out we don't have to know the functional form of the posterior $P(\theta|D)$ in order to simulate random samples from it.
- So, instead of summarizing $P(\theta|D)$, we summarize a very large number of samples from $P(\theta|D)$ as an approximation.
- We use those samples to perform inference.
- The two most popular MCMC algorithms for producing these samples are the Gibbs sampler and the Metropolis-Hastings sampler (MH).
- The packages in R for doing Bayesian inference will all revolve around using Gibbs and MH samplers.

What is a Gibbs sampler?

- It turns out if you know the full conditional $P(\beta|\sigma^2, X, Y)$ and $P(\sigma^2|\beta, X, Y)$, you can sample $P(\beta, \sigma^2|X, Y)$ like this:

$$
\begin{aligned}
&\text{Initialize } \beta_{(1)}, \sigma^2_{(1)} \\
&\text{For } t = 1 : T \\
&\qquad \beta_{(t+1)} \sim P(\beta_{(t)}|\sigma^2_{(t)}, X, Y) \\
&\qquad \sigma^2_{(t+1)} \sim P(\sigma^2_{(t)}|\beta_{(t+1)}, X, Y) \\
&\text{END}
\end{aligned}
$$

- After what is called a burn-in period, this algorithm generates samples from $P(\beta, \sigma^2|X, Y)$.
- If you know how to sample from both $P(\beta|\sigma^2, X, Y)$ and $P(\sigma^2|\beta, X, Y)$ directly, this is called a Gibbs sampler.
- Conjugate priors are so popular because we can use the Gibbs sampler, which is very fast to run, and easy to code.

What is M-H?

- What if you don't know either $P(\beta|\sigma^2, X, Y)$, $P(\sigma^2|\beta, X, Y)$, or both?
- Can you still sample from $P(\beta, \sigma^2|X, Y)$? Yes! Use Metropolis-Hastings.
- I don't want to get into all the details, but MH is very similar to Gibbs.
- Suppose I can't sample from $P(\beta|\sigma^2, X, Y)$.
- Proceed as if in a Gibbs sampler, except when it is time to sample $\beta_{(t+1)}$ do the following:
  - Generate β_* from a distribution we can sample from, we call this the proposal distribution, $p(\beta_*|\beta_{(t)})$.
  - Accept β_* and set $\beta_{(t+1)} = \beta_*$ with probability π_*, else reject β_* and set $\beta_{(t+1)} = \beta_{(t)}$.
  - Where $\pi_* = \min\left[\frac{P(\beta_*|X,Y,\sigma^2)p(\beta_{(t)}|\beta_*)}{P(\beta_{(t)}|X,Y,\sigma^2)p(\beta_*|\beta_{(t)})}, 1\right]$

R examples section 3

Why use Bayes?

- Ease of interpretation. Interpretation of the posterior probability as a measure of evidence.
  - Credible interval vs confidence interval.
  - Posterior probability vs p-values.
- Does not rely on assumption that sample is large.
  - Generalized linear models
  - Mixed models
  - Nonlinear models
- It lends itself well to the sequential nature of experimentation.
  - Prior knowledge + Data $\rightarrow$ Posterior knowledge
  - Old posterior knowledge + New data $\rightarrow$ New posterior knowledge

Resources

- Cran Task View on Bayesian Inference: `http://cran.r-project.org/web/views/Bayesian.html`
- Official library subject: R (computer program language)
- Virginia Tech's library has a number of ebooks on R available for free for Virginia Tech students, faculty, etc.
- Recommended texts:
  - "Bayesian Methods: A Social and Behavioral Sciences Approach, Second Edition", Jeff Gill, ISBN: 1584885629.
  - "Data Analysis Using Regression and Multilevel/Hierarchical Models", Gelman and Hill, ISBN: 052168689X.
  - "Bayesian Data Analysis, Second Edition", Gelman and Rubin, ISBN: 158488388X.
- And of course, LISA!