Module 1: Bayesian Regression Analysis

BIOS 6624: ADVANCED STATISTICAL METHODS AND ANALYSIS

NICHOLE CARLSON, PHD

Typical Class Day

1. New information (lecture and slides)
2. Working on an in-class analysis to practice data analysis and the concepts we are doing.
3. End with questions about the current project
4. Every other class work together on your current project together for 15-20 min.
5. There is an offset of the current project you have and the current material.

    - For example, we are doing Bayesian regression now, but you have Project 0 which doesn't require these skills. Part way through the module you will present your project and then have a project using the material from the current module. (Check in syllabus).

Goals of this module

1. Understand Bayesian thinking.
2. Understand how to conduct a regression analysis (confounding, mediation, interaction, model selection).
3. Be able to implement a Bayesian regression analysis.
4. Be able to communicate about your analysis orally and in writing.

So, we will be mixing learning about traditional regression topics and looking at them in a Bayesian framework.

Bayes' Rule

Bayes' Rule in Practice: Diagnostic Tests

Suppose 1 % of the population has a type of cancer (C) and a screening test for that cancer (T) has 80% sensitivity and 95% specificity.

Notation:

Probability of cancer: P(C)= 0.01

80% sensitivity: Pr(+Test|C) = 0.80

95% specificity: Pr(-Test|No C) = 1 - Pr(+Test|No C) = 0.95

What we really want to know is the probability I have cancer if I have a positive test or Pr (C|Test +).

Derive P(C|T+) using Bayes Theorem

Bayes' Rule: Diagnostic Tests

In general, people are mixing up

Pos. Pred. Value (PPV) = Pr(C|+Test) and

Sensitivity = Pr(+Test | C)

is analogous to the "Prosecutor's Fallacy" in legal settings - a small probability of evidence given innocence need not mean a small probability of innocence given evidence.

Another way to think about Bayesian:

Bayes' Rule gives a consistent way to update our (un)certainty in whether someone has C

With no other knowledge (*a priori),* what might you naturally conclude the probability a person has cancer is in this problem?

The prevalence: P(C) = 0.01.

Knowing the test results (+/-) leads to an *a posteriori* evaluation:

$$Pr(C|+Test) = 0.14 \text{ and } Pr(C|-Test) = 0.002$$

Bayes' Rule

.... **is more general, and the basis for the Bayesian approach to statistics.**

Using Bayes' Theorem, we have a way to update our a priori uncertainty in the value of model parameters (like the mean or a regression coefficient) in the light of the observed data.

Major elements of Bayesian Analysis

Let *Y* be the observed data and θ.

1. The model (or likelihood) of the data (e.g., a regression type model). L(Y| θ).

2. The prior: *A priori knowledge about the values of the parameters*. p(θ) or $\pi(\theta)$.

3. The posterior: *Knowledge about the values of the parameters given the data you observed.* $p(\theta|Y)$.

(1) and (2) are things that we will write down as part of developing our analysis and (3) is something we want to know.

Bayes Rule and the Bayesian Approach

Bayes' Rule allows us to get at what we want—updated information about θ given the data (Y).

Note: we normally write posterior $\propto$ likelihood x prior

An Example: Pain Relief Treatment

We perform a study of the effect of an analgesic. We wish to know the proportion that will find pain relief in the population (θ).

Step 1: Define a model for our data (d=the number of people finding pain relief) given a sample size (N=12 to start).

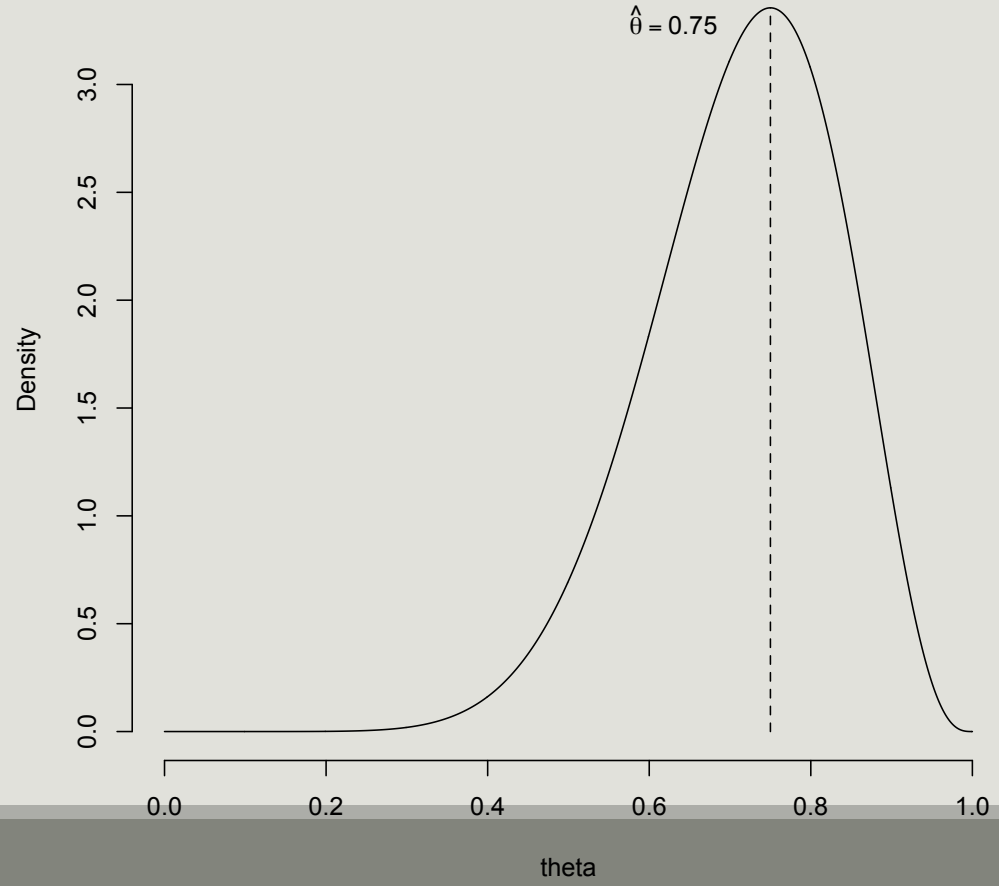
If d=9 find pain relief out of n=12 participants, what is the likelihood.

Frequentist “Usual Approach”: Maximum Likelihood and Conf. Intervals

The maximum likelihood estimator (MLE) of is value at which the likelihood is maximized for a given set of observed data.

If d=9 in our study, the MLE is the sample proportion 9/12 = 0.75.

Uncertainty in θ can be expressed as a confidence interval: (0.47,0.93).

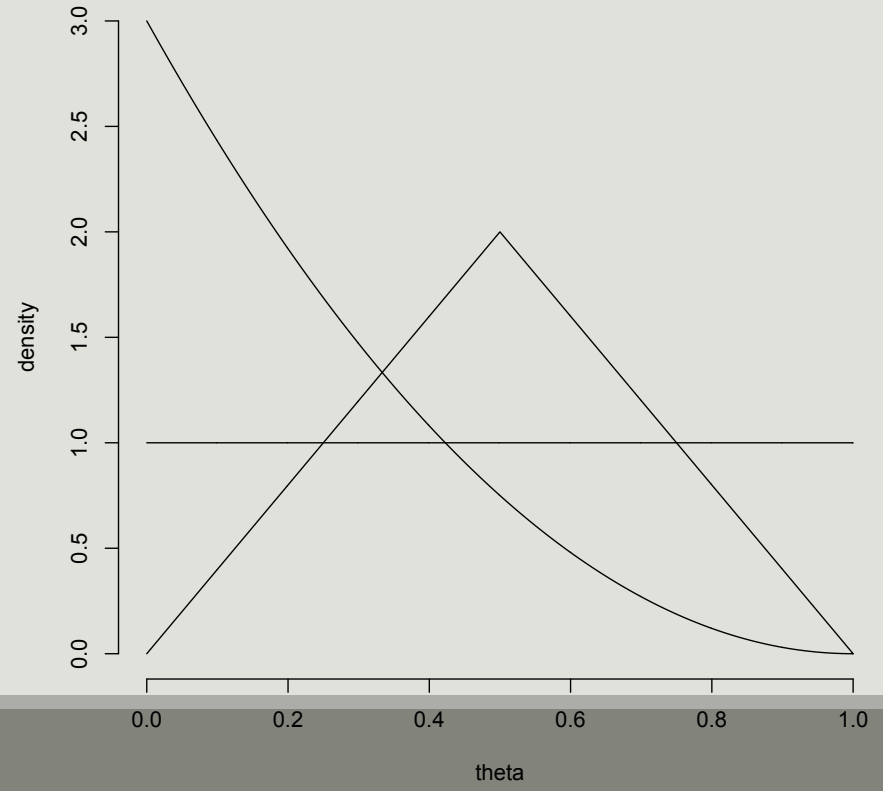


Bayesian Approach: The prior (you must specify this)

Bayesian analysis is to account for uncertainty in the unknown θ by specifying a prior distribution over [0,1], the range of possible values of θ.

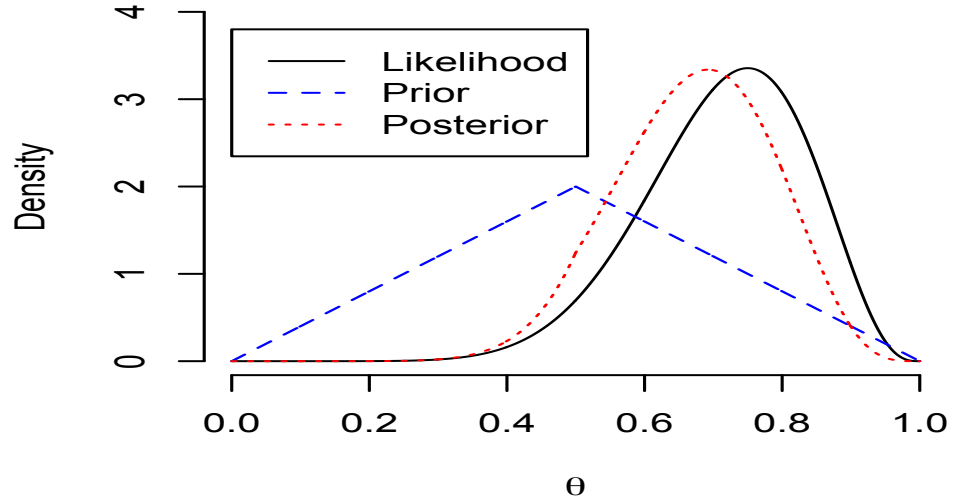
Note, the prior $p(\theta)$ expresses the experimenter's knowledge about θ excluding information in the data.

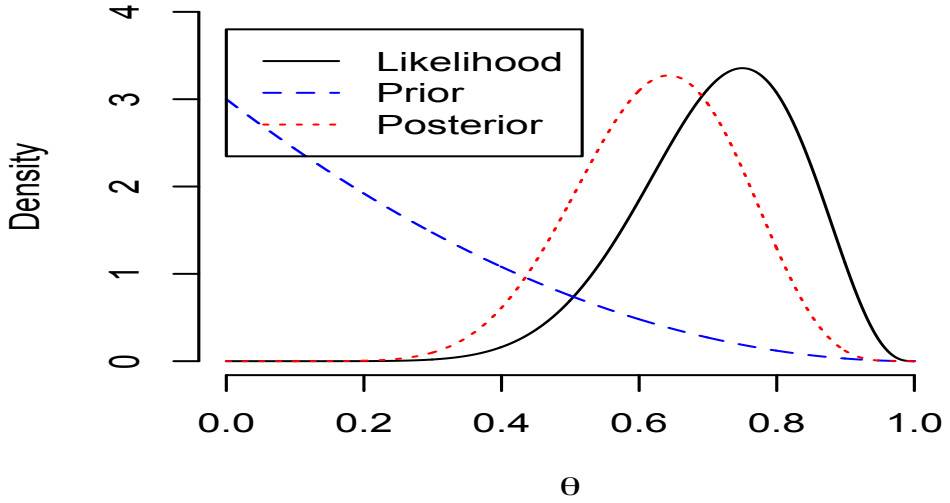
- A prior $p(\theta) = 1$ describes belief that all values of are equally likely a priori.
- The “triangular" prior gives more weight to values near 0.50.
- The Partial- U prior gives more weight to small values

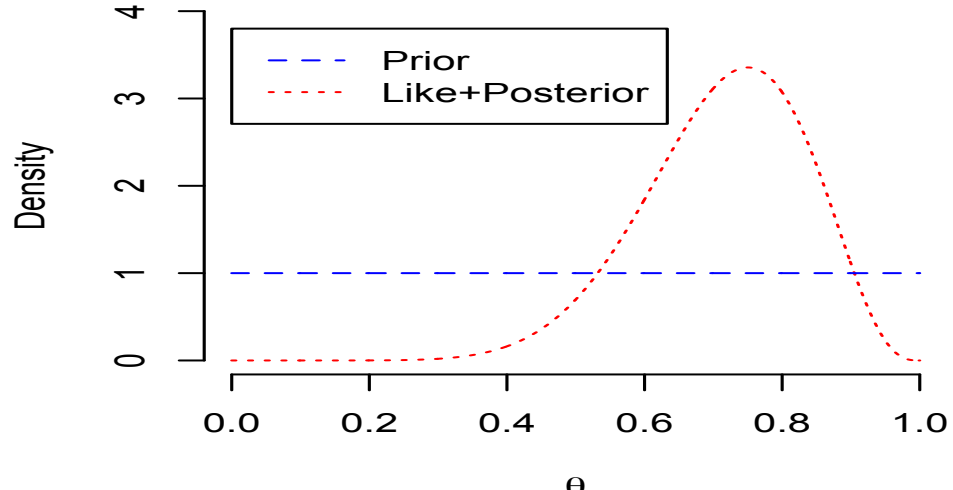


Prior	Mean	80% Prob. Interval
Flat	0.5	0.1-0.9
Triangle	0.5	0.22-0.78
Partial-U	0.25	0.03-0.53

Prior + Likelihood → Posterior: Analgesic Trial







Calculating the Posterior

In certain statistical models with standard (reference or conjugate) priors, the posterior distribution has a known distributional form (i.e. normal, beta, gamma) that is amenable to direct mathematical analysis and simple computations for posterior summaries.

1) Write down the likelihood.

2) Write down the prior.

3) Look for the core of a distribution (that is the posterior)

More generally, simulation methods are needed to compute the posterior distribution and posterior summaries.

- Modern computational methods based on Markov Chain Monte Carlo (MCMC) make Bayesian methods widely applicable and relatively easy to obtain (in SAS, Stata, BUGS).

Summarizing the Posterior: Analgesic Trial

- The mean (or median, or mode) of the posterior distribution of θ is a Bayes estimate of θ. (FYI: MLE is posterior mode with a flat prior.)

- Posterior probability (credible) interval: analog of confidence interval. Differences among posteriors for different priors reflect small sample size. With a uniform or flat prior, there is a 90% posterior probability that $0.55 \leq \theta \leq 0.85$.

Prior	Posterior Mean	90% Posterior Interval
Flat	0.71	0.55-0.85
Triangle	0.67	0.53-0.82
Left-U	0.63	0.47-0.77

- Smaller differences among posteriors with more data: if $d = 27$ and $n = 36$ (i.e. 3 times as many successes and failures):

Prior	Posterior Mean	90% Posterior Interval
Flat	0.74	0.64-0.82
Triangle	0.72	0.62-0.81
Left-U	0.70	0.61-0.79

Summary: Major elements of Bayesian Analysis

1. The model (or likelihood) of the data (e.g., a regression type model).

2. The prior: *A priori knowledge about the values of the parameters*.

3. The posterior: *Knowledge about the values of the parameters given the data you observed.*

Now let’s apply these concepts to a simple linear regression framework

Our In-class Project A Description of the project and questions of interest:

PSA is a well-established screening test for prostate cancer. A university medical center urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. The data consist of 97 men who were about to undergo radical prostectomies. The questions of interest for this study are:

1) What factors are associated with PSA levels?

2) Do the associations depend on whether there is seminal vesicle invasion?

Any questions for me as the investigator

Step 1 of any project….questions for the investigator.

Step 2: Reading in the Data and Descriptives

Follow through the worksheet to do Step 2