Introduction to Bayesian Methods

Arthur Small

Spring 2020

SYS 6014 Decision Analysis

Big picture: data-driven decision-making

You have a set of possible actions you can take.

You have a prediction model or classification model that enables you to *differentiate* between cases, based on data.

Your decision model then enables you to discriminate — to make the "right" intervention for each different case, rather than just choosing the same action for all cases.

Predictive models

Predictive models may take many different forms.

- They may use many, many different types of data
- They may be built using different analytic techniques: statistical regression, machine learning,...

We embrace this diversity of approaches.

But: we want our predictions to include *uncertainty* information.

So: want outputs in form of *probability distributions* over decision-state variables

Probabilistic predictions

Bayesian Methods

Reminder example: conditional probabilities

Example: Social mobility

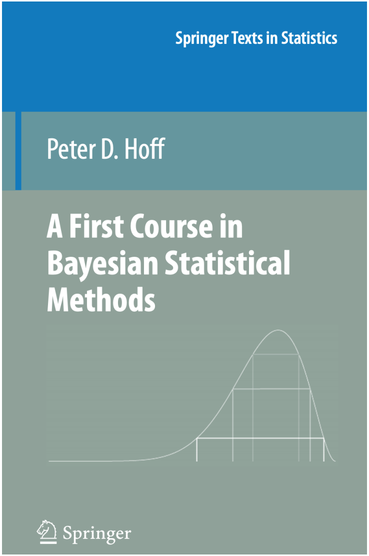
Logan (1983) reports the following joint distribution of occupational categories of fathers and sons:

	son's occupation				
father's occupation	farm	operatives	craftsmen	sales	professional
farm	0.018	0.035	0.031	0.008	0.018
operatives	0.002	0.112	0.064	0.032	0.069
craftsmen	0.001	0.066	0.094	0.032	0.084
sales	0.001	0.018	0.019	0.010	0.051
professional	0.001	0.029	0.032	0.043	0.130

Suppose we are to sample a father-son pair from this population. Let Y_1 be the father's occupation and Y_2 the son's occupation. Then

$$\begin{aligned}
\Pr(Y_2 = \text{professional} | Y_1 = \text{farm}) &= \frac{\Pr(Y_2 = \text{professional} \cap Y_1 = \text{farm})}{\Pr(Y_1 = \text{farm})} \\
&= \frac{.018}{.018 + .035 + .031 + .008 + .018} \\
&= .164\,.
\end{aligned}$$

Related readings: Hoff Chs. 1, 3



Bayesian methods: Introduction via simple example

Suppose you want to estimate the fraction of a population that is infected with some disease.

$\theta \in [0,1]$: true value

Test a random sample of 20 from the population.

$Y \in \{0,1,\ldots,20\}$: # of positive results.

Question: What does realized value of Y tell us about the true value of θ?

Sampling model

$Y|\theta \sim$ binomial$(20, \theta)$: For $y = 0, 1, \ldots, 20$, (i.i.d.)

$$l(y|\theta) = \Pr(Y = y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{(20-y)}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

$l(y|\theta)$ called the *likelihood function*.

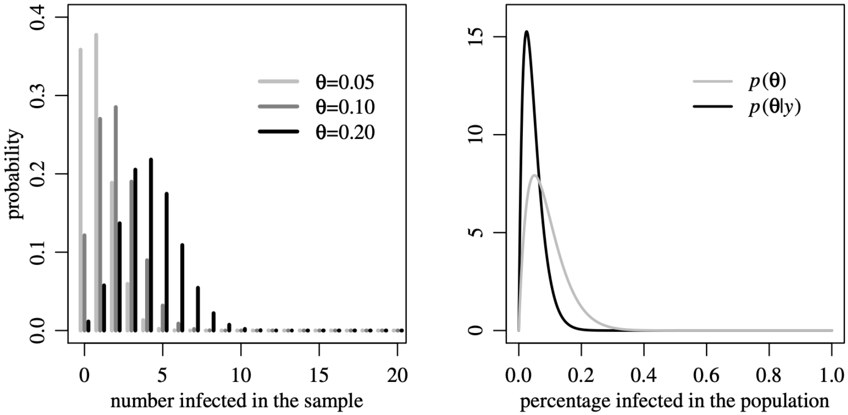


Fig. 1.1. Sampling model, prior and posterior distributions for the infection rate example. The plot on the left-hand side gives binomial$(20, \theta)$ distributions for three values of θ. The right-hand side gives prior (gray) and posterior (black) densities of θ.

Idea: For any $0 < \theta < 1$, all values of Y are *possible*, but some are more likely than others.

The likelihood function tells us how likely is each possible observation, for a given θ.

If, say, $Y = 15$, that provides evidence that θ is not small.

Core of Bayesian reasoning: work out all the different combinations of Y, θ that could have generated the observed sample data.

Prior information

Suppose we have some background knowledge about the likely values of θ.

Represent this knowledge by means of a *prior distribution* $\pi(\theta)$ over $[0, 1]$.

Obviously, there are many (infinitely many) possible such distributions.

For convenience, we typically choose to model priors as chosen from a parametrized family of distributions.

The Beta distribution

$$\theta \sim \mathrm{beta}(a, b)$$

Then

$$E[\theta] = \frac{a}{a+b}$$

For our case, let's suppose our prior beliefs correspond to:

$$\theta \sim \mathsf{beta}(2, 20)$$

$$\theta \sim \text{beta}(2, 20)$$

implies

$$\begin{aligned} \mathrm{E}[\theta] &= 0.09 \\ \mathrm{mode}[\theta] &= 0.05 \\ \Pr(\theta < 0.10) &= 0.64 \\ \Pr(0.05 < \theta < 0.20) &= 0.66\,. \end{aligned}$$

Bayes Theorem

Let $\pi(\theta|y)$ denote our *posterior distribution* over values of θ.

This means: our *updated* beliefs about the likelihood that θ takes various values, *after* we've received our test results.

Bayes Theorem says:

$$\pi(\theta|y) = \frac{l(y|\theta)\pi(\theta)}{Pr\{Y = y\}} = \frac{l(y|\theta)\pi(\theta)}{\int_{\Theta} l(y|\tilde{\theta})\pi(\tilde{\theta})d\tilde{\theta}}$$

Can be shown:

If $\theta \sim \text{beta}(2, 20)$ and $Y = 0$, then $\theta|y \sim \text{beta}(2, 40)$.

More generally:

If $\theta \sim \text{beta}(a, b)$ and $Y = y$, then $\theta|y \sim \text{beta}(a + y, b + 20 - y)$.

$$
\begin{aligned}
\mathrm{E}[\theta|Y=y] &= \frac{a+y}{a+b+n} \\
&= \frac{n}{a+b+n}\frac{y}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b} \\
&= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0,
\end{aligned}
$$

where $\theta_0 = a/(a+b)$ is the prior expectation of θ and $w = a+b$.

Sensitivity analysis

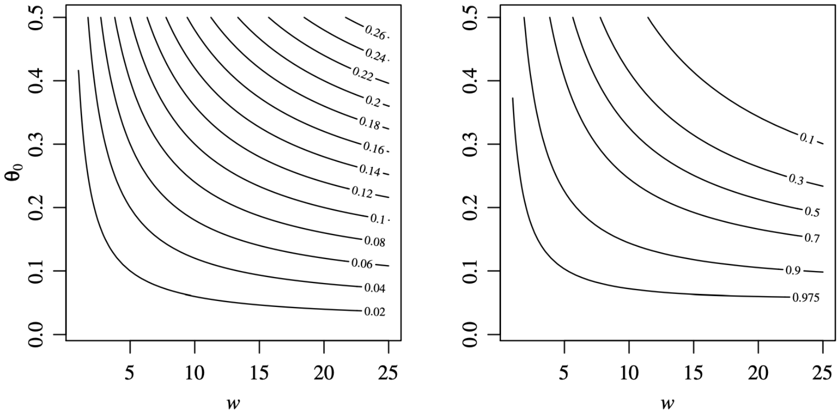


Fig. 1.2. Posterior quantities under different beta prior distributions. The left- and right-hand panels give contours of $\mathrm{E}[\theta|Y=0]$ and $\Pr(\theta < 0.10|Y=0)$, respectively, for a range of prior expectations and levels of confidence.

Building a predictive model

Sampling model and parameter space

Letting Y_i be the diabetes progression of subject i and $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,64})$ be the explanatory variables, we will consider linear regression models of the form

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{64} x_{i,64} + \sigma \epsilon_i.$$

The sixty-five unknown parameters in this model are the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{64})$ as well as σ, the standard deviation of the error term. The parameter space is 64-dimensional Euclidean space for $\boldsymbol{\beta}$ and the positive real line for σ.

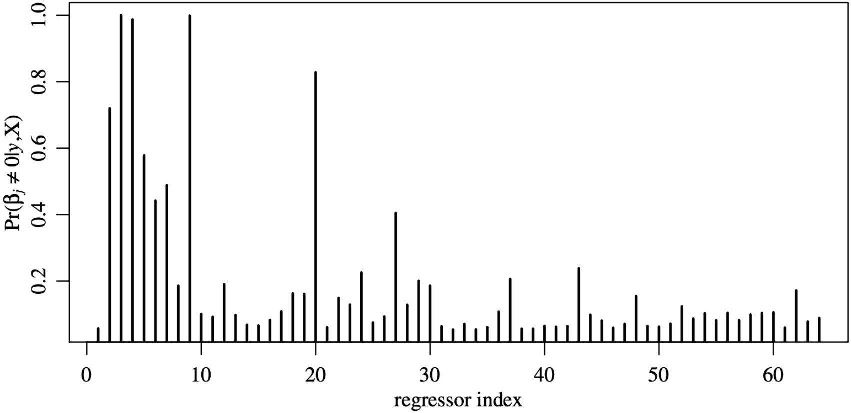


Fig. 1.3. Posterior probabilities that each coefficient is non-zero.

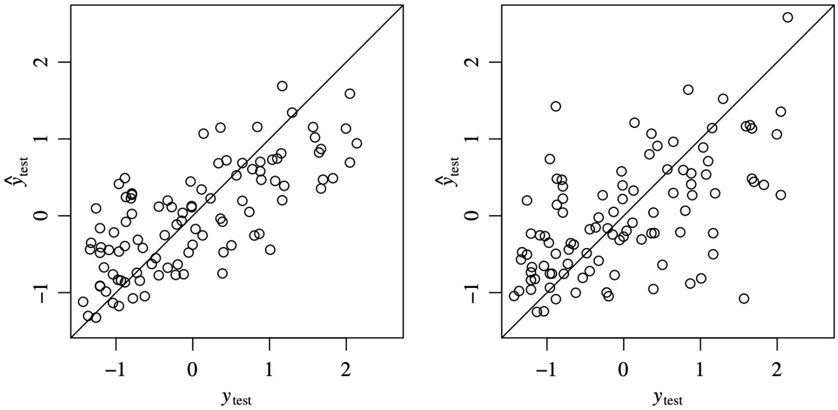


Fig. 1.4. Observed versus predicted diabetes progression values using the Bayes estimate (left panel) and the OLS estimate (right panel).

Prediction via the predictive distribution

$$
\begin{aligned}
\Pr(\tilde{Y} = 1 | y_1, \ldots, y_n) &= \int \Pr(\tilde{Y} = 1, \theta | y_1, \ldots, y_n) \; d\theta \\
&= \int \Pr(\tilde{Y} = 1 | \theta, y_1, \ldots, y_n) p(\theta | y_1, \ldots, y_n) \; d\theta \\
&= \int \theta p(\theta | y_1, \ldots, y_n) \; d\theta \\
&= \mathrm{E}[\theta | y_1, \ldots, y_n] = \frac{a + \sum_{i=1}^n y_i}{a + b + n} \\
\Pr(\tilde{Y} = 0 | y_1, \ldots, y_n) &= 1 - \mathrm{E}[\theta | y_1, \ldots, y_n] = \frac{b + \sum_{i=1}^n (1 - y_i)}{a + b + n} \,.
\end{aligned}
$$