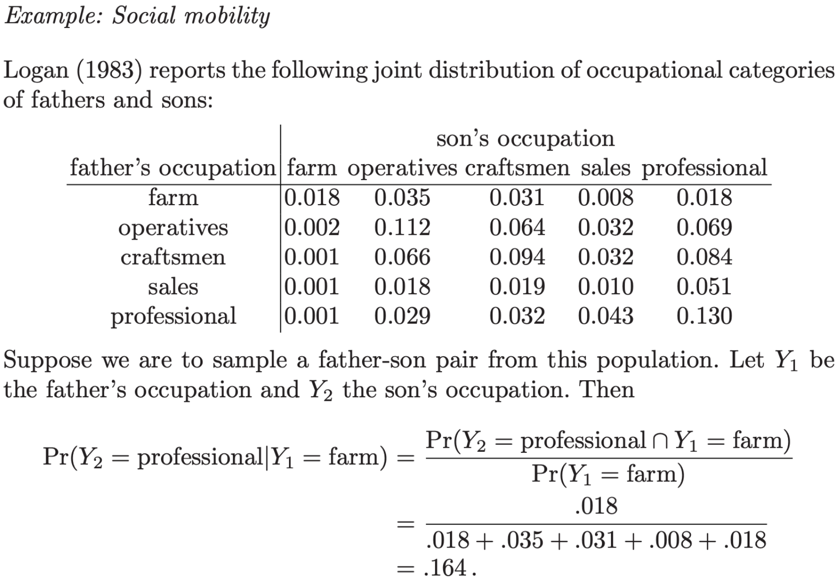
Session 11

Example: We have a population of fathers and their sons. Their occupation is divided into five categories. The numbers in the table show the joined probability distribution. Also, the sum of all numbers, in the table, equals to one.



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

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

Figure 1: 1

Two variables are independent if knowing one of them does not give you information about the other one. Also, two variables are independent if the following relationship is true between them:

$P(A \cap B) = P(A) \times P(B)$

What is the probability that the son is professional?

$Pr(Y_2 = Professional) = \sum Pr(Y_2 = Professional, Y_2 = i) = 0.352 \neq 0.164$

It is shown that the conditional probability is not equal to unconditional probability.

The above concept is essentially what your predictive model is going to do. You are going to go from unconditional probability to conditional probability. You are going to condition on some data you have to give the probability of some future event. There are different ways to create these joined probability events such as machine learning and so on. But essentially what you are going to do is to build somehow the joined probability distributions between the data that you know about (Prior knowledge) and the data you want to know about.

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.

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

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

Sampling model

$Y|\theta \sim \text{binomial}(20, \theta)$: For $y = 0, 1, \dots, 20$, (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*.

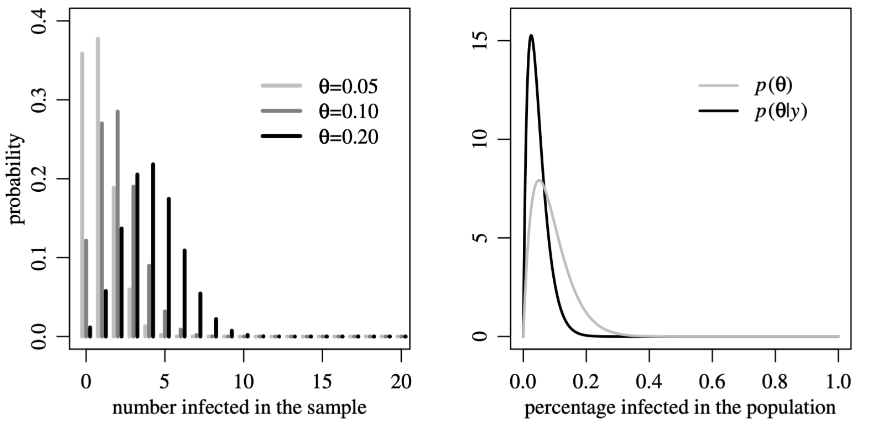


Fig. 1.1. Sampling model, prior and posterior distributions for the infection rate example. The plot on the left-hand side gives binomial$(20, \theta)$ distributions for three values of θ. The right-hand side gives prior (gray) and posterior (black) densities of θ.

Figure 2: 2

The above left graph is the graph of the likelihood function or the histogram of the likelihood function for different values of data. If θ is equal to 0.05, there is a substantial probability of getting zero successful tasks out of 20 and if you get zero success out of 20 that just not tell you the expected value for the real θ is zero.

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

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

Then

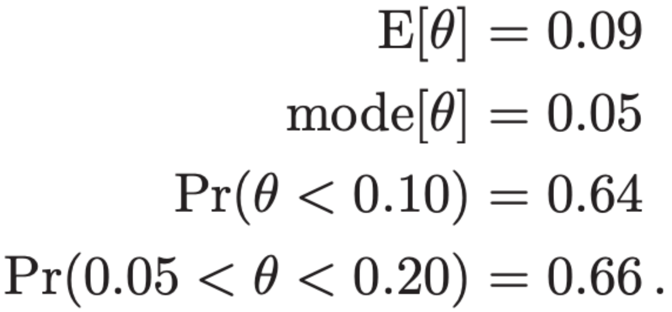
$$E[\theta] = \frac{a}{a+b}$$

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

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

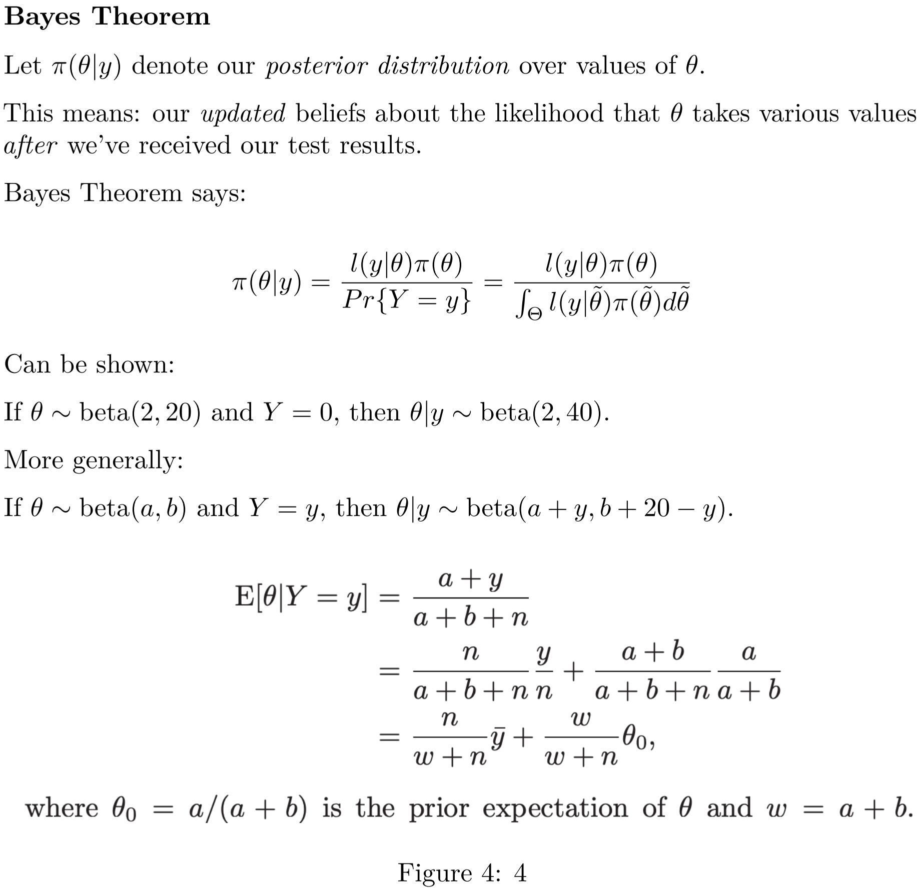
$$\theta \sim \text{beta}(2, 20)$$

implies



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

Figure 3: 3



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

More generally:

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

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

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

Figure 4: 4

The new expected value is being the weighted average of your prior expectation and the sample mean. If w is very small, then you are putting most of your weight on your sample data, otherwise you put most of the weight on the prior expectation.

Sensitivity analysis

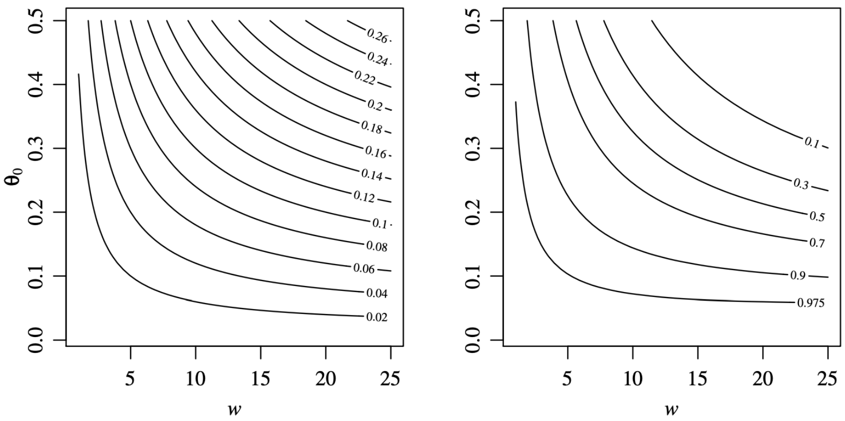


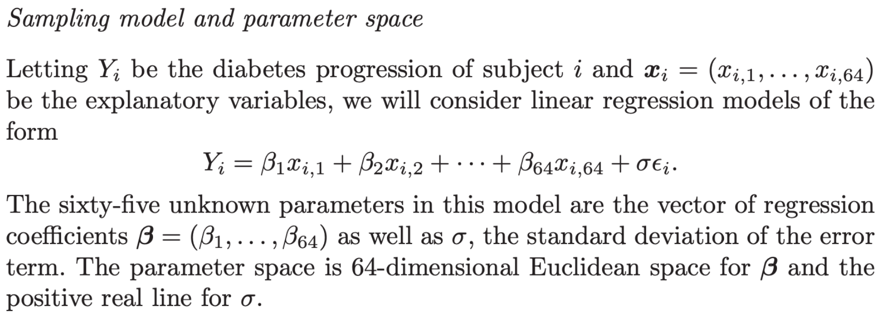
Fig. 1.2. Posterior quantities under different beta prior distributions. The left- and right-hand panels give contours of $\mathrm{E}[\theta|Y=0]$ and $\Pr(\theta < 0.10|Y=0)$, respectively, for a range of prior expectations and levels of confidence.

Figure 5: 5

The X-axis shows the w, and the Y-axis shows the prior expectation. Given the prior expectation is 0.5 and the number of the positive tests is equal to zero. If you have high strong prior beliefs, your posterior mean is going to be about 0.26. If you have less strong prior beliefs, your posterior mean is going to be about 0.16. If you have very weak strong prior beliefs, your posterior mean is going to be about 0.02.

Building a predictive model

Ex: We have a sample of 342 diabetes patients, and for each of them we have an observation for 64 different variables. These variables may or may not give you useful information. Suppose that the progress of diabetes is the linear function along with different variables. So, each wight indicates how important is the corresponding variable. We have an error process in this model, which is independent and identically distributed. The error follows the standard normal distribution. There is data dependence, such as the data dependence between family members who live at the same house. Most of these variables may not be useful and relevant.



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.

Figure 6: 6

Consider you find out the distribution of the θ given the sample expectation, now you want to figure out what is the likelihood of getting a different value for the next.

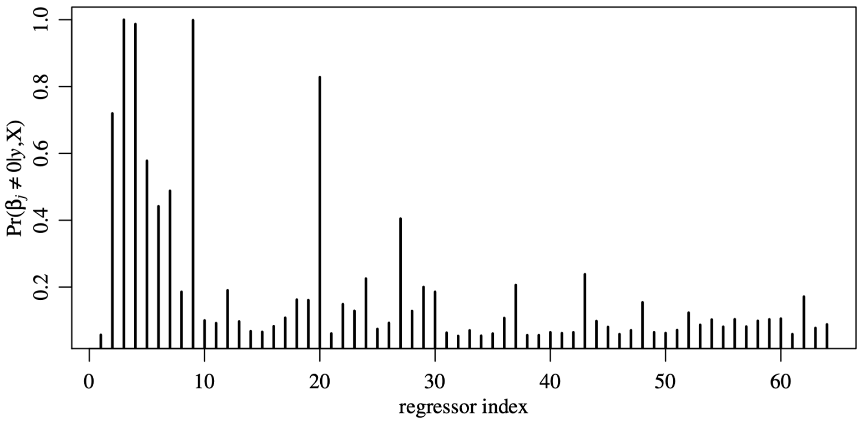


Fig. 1.3. Posterior probabilities that each coefficient is non-zero.

Figure 7: 7

In the above-left side graph, the predictions were plot against the observations. Here, if the model was perfect, every point will be exactly on the line.

To have a useful prediction, you should be cautious that it cannot depend on anything unobserved. You have to get rid of all of the unobserved parameters.

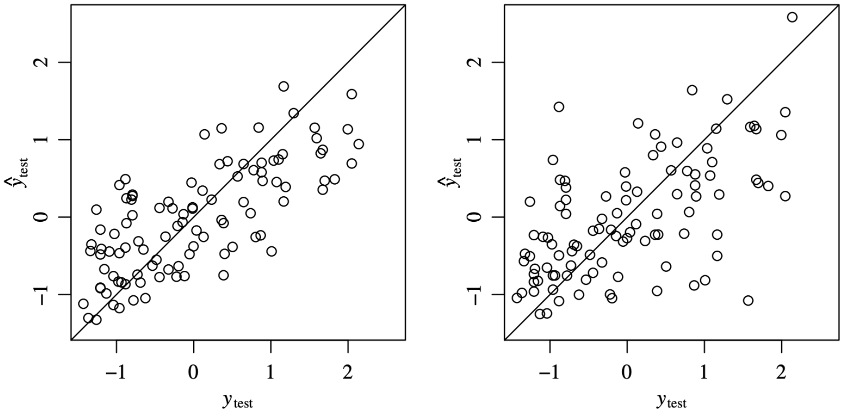
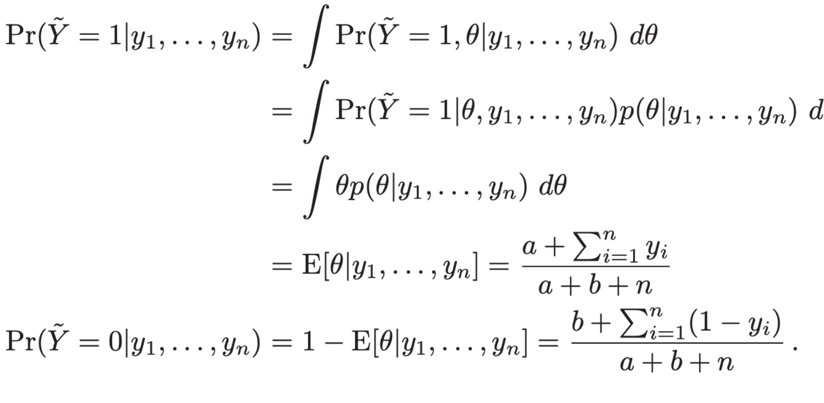


Fig. 1.4. Observed versus predicted diabetes progression values using the Bayes estimate (left panel) and the OLS estimate (right panel).

Figure 8: 8



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

Figure 9: 9