STA 250: Statistics

Notes 8. Inference with Bayes: Prediction

Book chapters: (11.1-2 relate to the last part)

1 Predict instead of testing

Example (TC count trend.)**.** Recall the annual TC count example where we model TC counts $X = (X_1, \cdots, X_n)$ from last $n = 100$ as $X_t \overset{\text{IND}}{\sim} \mathsf{Poisson}(\alpha\beta^{t-1})$, $t = 1, \cdots, n$, $\alpha > 0, \beta > 0$. An assessment of whether TC counts are trending could be done by looking at various possible statements about β, such as "$\beta > 0$", "$\beta = 0$" etc and judging their plausibility given data. But β is not exactly a tangible physical quantity, it's rather an artifact of the model assumed. So from the point of scientific analysis, it could be more interesting to assess what is going to happen to TC counts in the coming years. For example we could look at X^* =TC count on the 10th year into the future and try to make predictions about X^* based on data on X. The model in this case serves just as a bridge between X and X^*.

Example (NSW and income)**.** As another example, consider the NSW study where data $X = (X_1, \cdots, X_n)$ are available from $n = 185$ on the difference in their annual earnings before and after a job training program. Consider the model $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty), \sigma > 0$. Earlier we looked at evaluating various statements on μ, e.g., "$\mu \leq 0$", "$\mu = 0$", etc. But to assess the effectiveness of the program it could be more useful to evaluate what difference X^* it will make to the earning of a person enrolling in the future. Reporting a range for only μ does not answer this question. X^* has a variation σ around μ, and so predicting X^* must combine this variation with our uncertainty in μ.

In this context, direct inference on μ does have a prediction like interpretation. If you considered a large number of people enrolling in the program in the future, then their average earning difference can be equated with μ.

2 Posterior predictive of X^* given $X = x$

So our prediction problem is as follows: we have data X modeled as $X \sim f(x|\theta), \theta \in \Theta$ and we want predict a future observable quantity X^* based on actual recorded data $X = x$. To be able to make progress on this, we first need to link X^* to X through the parameter θ. In most cases, we will be able to write that given θ

$$X^* \sim f^*(x^*|\theta),\ x^* \in S^*,$$

for some pdf/pmf $f^*(x^*|\theta)$ over some set S^*, and take X^* to be conditionally independent of X given θ is specified [In the betting language, our bets on X^* given information on θ will not alter if we were given additional information on X].

Example (TC count (contd.))**.** For the TC count example, by identifying $X^* = X_{n+10}$ we may write $X^* \sim \mathsf{Poisson}(\alpha\beta^{n+10-1})$, and take X, X^* independent of each other given α, β.

Example (NSW (contd.))**.** For the NSW study, we could write $X^* \sim \mathsf{Normal}(\mu, \sigma^2)$ and again treat it to be independent of X given information on μ, σ.

The above formulation allows to write the joint pd/mf of (X^*, X) given θ as $f^*(x^*|\theta)f(x|\theta)$, $(x^*, x) \in S^* \times S$. With a prior pdf/pmf $\xi(\theta)$ on θ, we can now give a joint pd/mf of (X^*, X, θ) by the function

$$g(x^*, x, \theta) = f^*(x^*|\theta)f(x|\theta)\xi(\theta), \quad x \in S, x^* \in S^*, \theta \in \Theta.$$

Now let's factor in the observed data $X = x$ and look at the conditional description of the unobserved quantities (X^*, θ) given $X = x$. This must be given by the pd/mf

$$\begin{aligned} h^*(x^*, \theta|x) &= \frac{g(x, x^*, \theta)}{\text{const [to make into a pd/mf in } (x^*, \theta)]} \\ &= \frac{f^*(x^*|\theta)f(x|\theta)\xi(\theta)}{\text{const [to make into a pd/mf in } (x^*, \theta)]} \\ &= \frac{f^*(x^*|\theta)\xi(\theta|x)}{\text{const [to make into a pd/mf in } (x^*, \theta)]} \end{aligned}$$

because $f(x|\theta)\xi(\theta)$ and $\xi(\theta|x)$ are constant multiples of each other. Now the function $f^*(x^*|\theta)\xi(\theta|x)$ is indeed a pd/mf in (x^*, θ), so the normalizing constant in the last expression is simply 1 and we have

$$h^*(x^*, \theta|x) = f(x^*|\theta)\xi(\theta|x), \quad x^* \in S^*, \theta \in \Theta.$$

But we are interested in talking about only X^* given $X = x$. So we need to extract the marginal pdf/pmf of X^* from the above joint pd/mf, which gives us the "posterior predictive" pdf/pmf of X^* given $X = x$ in the function

$$f^*(x^*|x) = \begin{cases} \sum_{\theta \in \Theta} f^*(x^*|\theta)\xi(\theta|x) & \text{if } \xi(\theta|x) \text{ is a pmf} \\ \int_\Theta f^*(x^*|\theta)\xi(\theta|x)d\theta & \text{if } \xi(\theta|x) \text{ is a pdf} \end{cases}$$

Intuitively, the predictive distribution $f^*(x^*|x)$ stands for the following. If we knew θ, we would use $f^*(x^*|\theta)$ to describe X^*. But we do not know θ and our understanding of it is represented by the posterior pdf $\xi(\theta|x)$ given $X = x$. So we must combine our desription of X^* given θ with our current understanding of θ given $X = x$ to get $f^*(x^*|x) = \int_\Theta f^*(x^*|\theta)\xi(\theta|x)d\theta$.

3 Carrying out Bayesian prediction

So given data $X = x$, our understanding of X^* is completely captured in the posterior predictive pdf/pmf $f^*(x^*|x)$. Making prediction about X^* is equivalent to producing various

summaries of this pdf/pmf. For example to evaluate $X^* \in A$ for a given set of interest A, we would calculate

$$P(X^* \in A|X = x) = \begin{cases} \sum_{x^* \in A} f^*(x^*|x) & \text{if } f^*(x^*|x) \text{ is a pmf} \\ \int_A f^*(x^*|x)dx^* & \text{if } f^*(x^*|x) \text{ is a pdf.} \end{cases}$$

Also to produce a range for X^*, say a 95% range, we would extract the 2.5% and 97.5% quantiles of $f^*(x^*|x)$ and report the corresponding interval.

We can perform any of the above calculations if we have an easy formula for $f^*(x^*|x)$ to work with. But it is not necessary. Because of the special form of $f^*(x^*|x)$ we could in fact write: $P(X^* \in A|X = x) = \int P(X^* \in A|\theta)\xi(\theta|x)d\theta$ or $\sum_{\theta\in\Theta} P(X^* \in A|\theta)\xi(\theta|x)$ as appropriate. This will be helpful if we could quickly calculate $P(X^* \in A|\theta)$ for any θ, which is usually the case, as in the following example.

Example (NSW (contd.))**.** Consider a simplified model where $\sigma = 9000$ is assumed, and the unknown parameter $\mu \in \{-6000, -4000, -2000, 0, 2000, 4000, 6000\}$, with a discrete uniform prior $\xi(\mu)$ assigned to μ. Suppose we want to evaluate $X^* > 2000$ for recorded data X with $\bar{X} = 4253.57$. We know $P(X^* > 2000|\mu, \sigma) = 1 - \Phi(\frac{2000-\mu}{\sigma})$. And so,

$$P(X^* > 2000|X = x) = \sum_{\mu\in\{-6000,-4000,-2000,0,2000,4000,6000\}} \left[1 - \Phi\left(\frac{2000-\mu}{\sigma}\right)\right] \xi(\mu|x).$$

To calculate $\xi(\mu|x)$, recall that for the normal model $\ell_x(\mu) = \text{const} - n(\bar{x} - \mu)^2/\{2\sigma^2\}$. And hence we may write:

μ	$L_x(\mu) = \text{const.} \times \cdots$	$\xi(\mu)$	$\xi(\mu\|x) = \text{const} \times \cdots$	$\xi(\mu\|x)$	$1 - \Phi(\frac{2000-\mu}{\sigma})$
-6000	0	1/7	0	0	0.187
-4000	0	1/7	0	0	0.252
-2000	0	1/7	0	0	0.328
0	0	1/7	0	0	0.412
2000	0.003	1/7	0.003	0.003	0.500
4000	0.929	1/7	0.929	0.965	0.588
6000	0.031	1/7	0.031	0.032	0.672
			total = 0.963	1	

Therefore, $P(X^* > 2000|X = x) = 0.003 \times 0.500 + 0.965 \times 0.588 + 0.032 \times 0.672 = 0.59$.

4 Sampling based calculation

It might seem that to produce a 95% range of X^* we must get hold of the pdf/pmf $f^*(x^*|x)$. Not really. We could approximate this range, i.e., approximate the relevant quantiles, if we could draw a large sample of X^* values from $f^*(x^*|x)$. This is fairly easy to do once we go back to the joint density $h^*(x^*, \theta|x) = f^*(x^*|\theta)\xi(\theta|x)$ and recognize that we can draw an (x^*, θ) from the joint pd/mf by: first drawing a θ from $\xi(\theta|x)$, and then drawing an x^* from $f^*(x^*|\theta)$ with the same θ. We can repeat this many many times to gather a large number of draws $(x^{*(m)}, \theta^{(m)})$, $m = 1, \cdots, M$ (typically $M =$ a few hundreds or thousands). Then we ignore the θ part of the draws and just look at $x^{*(m)}$, $m = 1, \cdots, M$ which must be draws from $f^*(x^*|x)$.

Example (Opinion poll)**.** Suppose a researcher, having found $X = 200$ supporters of a policy among $n = 500$ students she surveyed, wants to predict the number of supporters X^* in another group of $m = 10$ students. Suppose p, the actual proportion of supporters in the college is assigned a $\mathsf{Uniform}(0, 1)$ prior. Then from the discussion above, we could get draws $x_1^*, \cdots, x_M^*$ from $f^*(x^*|x)$ as follows:

```
n <- 500; x <- 200;              ## data
a <- 1; b <- 1;                  ## prior Be(a = 1, b = 1)
a.x <- x + a; b.x <- n - x + b; ## posterior Be(a.x, b.x)
M <- 1000                        ## number of samples to draw
p.samp <- rbeta(M, a.x, b.x)     ## draw p from posterior
x.star <- rbinom(M, 10, p.samp) ## draw x.star[i] ~ Bin(10, p.samp[i])
```

The first six paired draws of p and x^* are:

```
> cbind(p.samp, x.star)[1:10,]
      p.samp x.star
 [1,]   0.42      4
 [2,]   0.41      4
 [3,]   0.40      1
 [4,]   0.43      6
 [5,]   0.39      2
 [6,]   0.43      5
```

To make prediction summaries, we ignore `p.samp` and just focus on summarizing `x.star`. Below are several graphical and numerical summaries.

```
> hist(x.star, freq = FALSE, col = "gray", border = "white", breaks = 0:11 - 0.5)
> ## get 95\% range
> quantile(x.star, c(.025, .975))
 2.5% 97.5%
    1     7
> ## calculate P(X* > 7 | X = x)
> mean(x.star > 7)
[1] 0.011
```