Stat 305A: Linear Models **Autumn 2019/20**

Lecture 17: More Bayes and random effects

Lecturer: Art B. Owen *November 21*

17.1 Bayes for regression

We looked at how Bayesian linear regression models work, referring to an example in Chapter 9 of Peter Hoff's book.

There was a linear model relating changed O_2 uptake as a function of age and another variable indicating whether the subject was on a running program or some other aerobic program. The data were differences so there was no ANCOVA element in this that one might otherwise have considered.

The largees model considered was

$$Y_i = \beta_1 + \beta_2 \mathrm{AER}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{AER}_i \times \mathrm{AGE}_i + \varepsilon_i.$$

which plots as two not necesarily parallel lines for $\mathbb{E}(Y)$ versus age. Here AER_i is 1 for subjects in the aerobic program and 0 for those who merely ran.

The likelihood for β with σ fixed is the exponential of a quadratic. If the prior has the same structure, then so does the posterior and we get a Gaussian posterior distribution for β so long as the quadratic inside the exponential is negative definite.

Let the Gaussian prior be

$$\beta \sim \mathcal{N}(\beta_0, \Sigma_0).$$

Then we get (see Hoff for the steps)

$$\begin{aligned} \mathrm{var}(\beta \,|\, y, x, \sigma^2) &= (\Sigma_0^{-1} + Z^\mathsf{T} Z/\sigma^2)^{-1} \quad \text{and} \\ \mathbb{E}(\beta \,|\, y, x, \sigma^2) &= (\Sigma_0^{-1} + Z^\mathsf{T} Z/\sigma^2)^{-1}(\Sigma_0^{-1}\beta_0 + Z^\mathsf{T} y/\sigma^2), \end{aligned}$$

where $x_i = (\mathrm{AGE}_i, \mathrm{AER}_i)^\mathsf{T}$ and $Z \in \mathbb{R}^{n\times 4}$ is our design matrix (observations $\times$ features), y is all the y_i and x is all the x_i. In class, we discussed how this looks the same as in the scalar case: inverse variances get summed and the means get weighted.

We need a prior for σ^2. Taking $\gamma = 1/\sigma^2$, we have

$$\gamma \sim \frac{\mathrm{Gam}(\nu_0/2)}{\nu_0\sigma_0^2/2}.$$

We will see in the posterior formulas that this prior is like having ν_0 prior observations (or degrees of freedom so maybe $\nu_0 + 1$ observations) with a sample variance of σ_0^2. Or,

$$\sigma^2 \sim \frac{\nu_0\sigma_0^2/2}{\mathrm{Gam}(\nu_0/2)}.$$

The posterior distribution for γ is

$$p(\gamma \,|\, y, Z\beta) \propto \gamma^{(\nu_0+n)/2-1} e^{-\gamma[\nu_0\sigma_0^2+SSR)(\beta)]/2}$$

which we recognize as

$$\gamma \sim \frac{\mathrm{Gam}((\nu_0+n)/2)}{(\nu_0\sigma_0^2+\mathrm{SSR})/2}$$

for $\mathrm{SSR} = \sum_i (y_i - z_i^\mathsf{T}\hat{\beta})^2$.

We do not dwell here on how to compute this posterior distribution. In passing we note that this case could be sampled by alternately sampling β given γ and γ given β. This is known as the ***Gibbs sampler***. There are many much more sophisticated ways to sample from a posterior distribution. Take a course in Monte Carlo or applied Bayes to see them. If you had a huge sample of (β, σ^2) pairs sampled from the posterior distribution, you could use them to estimate posterior means, variances, covariances and probabilities of interest to you.

17.2 Choosing β_0 and Σ_0

Ridge regression that penalizes the intercept too, is like taking $\beta \sim \mathcal{N}(0, \tau^2 I)$. If we don't want to regularize the intercept to be close to zero, we could replace $\Sigma_0 = \tau^2 I$ by

$$\Sigma_0 = \begin{pmatrix} M & 0 \\ 0 & I \end{pmatrix} \tau^2$$

for some large scalar value M. That allows the intercept to be large without incurring a prior penalty. Equivalently

$$\Sigma_0^{-1} = \begin{pmatrix} \epsilon & 0 \\ 0 & I \end{pmatrix} \tau^{-2}.$$

A flat non-informative prior with $M \to \infty$ or $\epsilon \to 0$ is close to what we do in ridge regression.

There is a unit information prior due to Kass and Wasserman. It takes

$$\Sigma_0^{-1} = \frac{Z^\mathsf{T}Z}{n\sigma^2} \quad \text{and} \quad \beta_0 = \hat{\beta}_{\mathrm{OLS}}.$$

It is then like a prior with the strength of just $n = 1$ prior observation right at the ordinary least squares estimate $\hat{\beta}_{\mathrm{OLS}}$. The practice of plugging some sample values into the prior is a form of ***empirical Bayes***. They also take $\nu_0 = 1$ and $\sigma_0^2 = \hat{\sigma}^2 = s^2$.

That prior is a special case of the Zellner prior

$$\beta \sim \mathcal{N}(\beta_0, g(Z^\mathsf{T}Z)^{-1}\sigma^2).$$

Letting $q = g/(g+1)$ (and quoting Wikipedia!), the posterior is

$$\beta \sim \mathcal{N}(q\hat{\beta}_{\mathrm{OLS}} + (1-q)\hat{\beta}_0, q(Z^\mathsf{T}Z)^{-1}\sigma^2).$$

For $\beta_0 = \hat{\beta}_{\mathrm{OLS}}$,

$$\beta \sim \mathcal{N}(\hat{\beta}_{\mathrm{OLS}}, q(Z^\mathsf{T}Z)^{-1}\sigma^2).$$

For $\beta_0 = 0$,

$$\beta \sim \mathcal{N}(q\hat{\beta}_{\mathrm{OLS}}, q(Z^\mathsf{T}Z)^{-1}\sigma^2).$$

and the posteror means has shrunk each component of $\hat{\beta}_{\mathrm{OLS}}$ by the factor $q = g/(g+1) < 1$. The Zellner posterior for $\gamma = 1/\sigma^2$ is

$$\frac{1}{\sigma^2} \,|\, x, y \sim \frac{\mathrm{Gam}((\nu_0 + n)/2)}{(\nu_0\sigma_0^2 + \mathrm{SSR}_g)/2}$$

where $\mathrm{SSR}_g = y^{\mathsf{T}}(I - qH)y$ and H is the usual hat matrix.

17.3 O_2 uptake example

He considers 5 models,

$$M \in \{\text{intercept only, AER only, AGE only, parallel lines, two lines}\}.$$

There could have been 8 models but these models avoid putting in an interaction AGE $\times$ AER unless ***both*** of those are also in the model. It is common to impose a ***hierarchy constraint*** like this on models. If an interaction of some order is present, then so are all sub-interactions. If a polynomial term is present then so are all lower order polynomial terms. It is not a theorem that this is always better, it is just a commonly adopted guideline.

Putting a uniform distribution on M and then for each model a Zellner prior on its parameters, generates a posterior distribution on β and σ^2. Any time that the AER only model is chosen we automatically have $\beta_3 = 0$ for the AGE coefficient. The posterior distribution of each β_j might then have a lump at zero and some other distribution for nonzero values.

The estimated posterior probabilities on those models are

$M =$	{intercept only,	AER only,	AGE only,	parallel lines,	two lines}
w prob.	(0.00,	0.00	0.18	0.63	0.19).

If we sample β or $z_{n+1}^{\mathsf{T}}\beta$ from this model (for fixed z_{n+1} representing a potential $n+1$'s subject) we will get predictions that do ***Bayesian model averaging*** instead of just picking the one possibly best model and using it alone. It is clear that AGE makes a difference, and less conclusive about AER. If we just want to predict, we don't have to decide which mdoel to go with. Believing the posterior there is at best 63% chance that we would choose correctly.

With a larger data set we would probably have the posterior probability just pick out one of these models. I would bet on the most general one here. This may be why Hoff chose an example with a small data set. It is more interesting.

He has another diabetes example with 64 predictors potentially $2^{64} \approx 10^{19}$ models (if you don't enforce hierarchy). Only 38 distinct models turned up in the 10,000 samples from the posterior distribution. Only 4 models turned up more than 6 times.

In that bigger example Bayesian model averaging was more accurate on held out data (MSE 0.452) than was OLS (MSE 0.67) or backward elimination (MSE 0.53). (I wonder how ridge or a Zellner prior would have done.)

17.4 Hierarchical models

Let $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Our plain ANOVA approach would have us choose between H_0 where $\mu_1 = \mu_2 = \cdots = \mu_k$ and H_A where the μ_i are completely unrelated to each other. In a

Bayesian approach, we can make $\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_*, \sigma_*^2)$. A tiny σ_* is like imposing H_0 and an enormous σ_* is like considering the μ_i to be unrelated. If k is large we should be able to learn something about μ_* and σ_* and find a middle ground between H_0 and H_A.

That middle ground could be very helpful. If $n_1 = 3$ and $n_2 = 1400$ then data from samples 2 through k help us learn μ_* and σ_* and shrink $\bar{Y}_{1\bullet}$ towards values from the other groups. We might find that we have enough data from group 2 that we don't need to shrink it's mean very much at all.

There can also be a hierarchical model for the σ_i.

If we pick a prior for the (μ_i, σ_i) pairs it will ordinarily have parameters. We can place a prior on those parameters. It could be a very flat one so that we then learn the extent to which data from one group are relevant to the others. We would ordinarily just pick one flat prior there instead of picking a prior that had further random parameters.

17.5 Spike and slab

We might believe that our vector β has lots of tiny values and a small number of meaningfully large values. We can model that by taking a prior with independent

$$\beta_j \sim \lambda\mathcal{N}(0, \epsilon^2) + (1-\lambda)\mathcal{N}(0, M^2)$$

where ϵ is tiny and M is large and $0 < \lambda < 1$ is a mixture parameter. The first component is a spike near zero and the second is a slab over a wide swath of values.

The normal mixture components could be replaced by double exponential. The spike could be a point mass at zero, i.e., $\mathcal{N}(0, 0)$. The slab could be improper.

The posterior distribution of β_j could well be bimodal, sometimes near zero and sometimes near some other value.

17.6 Bayesian software and examples

In this course, we have just scratched the surface of Bayes.

Real world applications commonly take you way beyond what can be done with closed forms and conjugate distributions.

There are probabilistic programming languages that let you specify what your prior is, and where your data are, and they take it from there. I would recommend STAN as a start at least for modest sized data sets and hierarchical models.

At present there is always the possibility that your Bayesian software fails to properly sample from the posterior distribution you have in mind. There are diagnostics. Most of them can be fooled.

If you are looking for detailed worked examples, there is a Stan con conference that includes many of them.

17.7 Random effects

The random effects model is a non-Bayesian counterpart to hierarchical Bayesian models. We looked at how it plays out in the two factor setting. You could have two fixed effects, two random effects, or one of each. The fixed $\times$ random effects setting is the one where you can most easily get it wrong in a consequential way. See Chapter 11 of the notes by Eric Min. Those notes outline the sums of squares and mean squares and F tests.

If you have n observations on k people (random effects) for t treatments (fixed effects) then while you have $n \times k \times t$ data values you don't really have as much information as that might make you think. To see why, let $n \to \infty$ and take means per person. Then you have data about t treatments for k people and your sample size is obviously just k vectors in $\mathbb{R}^t$ if you're thinking about what you've learned about people. You're ordinarily better off measuring $2k$ people n times each than measuring k people $2n$ times each. Getting $2k$ people is probably going to cost more to do. After thinking about it this way, the initially strange looking F-tests for crossed random effects may make more sense.

Whether an effect is random or fixed depends in part on our goals. If there are only k levels of some effect then it is fixed. If instead we have data on k levels and we want to generalize to a universe of $K \gg k$ levels from which they were sampled, then the effect is a random effect. If there are $K \gg k$ levels but we only want to think about the k we have studied, then it is back to a fixed effect analysis. E.g, there are k people who run the lathes in our machine shop. For differences among those k people a fixed effects analysis is appropriate. To draw conclusions on equipment operators in general, of whom we just happen to have k examples, then it is a random effect.