Econometrics, Lecture 17. Bayesian Statistics

Ron Smith
EMS, Birkbeck, University of London

Autumn 2020

Last time

- We discussed the ways exogeneity could fail and the implications for OLS
- Looked at the identification problem and the the simultaneous equations model
- Looked at instrumental variable/two stage least squares estimates
- Now consider Bayesian statistics, distinguish
  - true believers: committed Bayesians
  - pragmatic Bayesians who only use it for particular problems where frequentist methods do not work, such as. Bayesian Model Selection, BIC; Bayesian DSGEs; and Bayesian VARS; and
  - ad hoc Bayesians who use priors and Bayesian interpretations when doing frequentist statistics.

Contrasting approaches

Frequentist statistics

- Probabilities are the limits of relative frequencies as the sample size goes to infinity
- Parameters as fixed numbers;
- Imagine sampling distribution over lots of hypothetical samples of which the data is just one;
- Uses the Neyman-Pearson hypothesis testing framework

Bayesian Statistics

- Probabilities measure degrees of belief
- Parameters are random variables
- Uses prior distributions for the parameters based on past experience and uses Bayes rule to provides a systematic way to update beliefs
- Estimation and inference.is done conditional on the observed data not sets of hypothetical samples
- Has explicit loss functions based on decision criteria and make decisions rather than doing tests.

Bayes Theorem

- Bayes Theorem for continuous variables A and B follows from the definitions of conditional probability in terms of the joint, $f(A, B)$, and marginal probabilities

$$f(A \mid B) = \frac{f(A, B)}{f(B)}$$

$$f(B \mid A) = \frac{f(A, B)}{f(A)}$$

$$f(A, B) = f(B \mid A) f(A)$$

- giving Bayes Theorem

$$f(A \mid B) = \frac{f(B \mid A) f(A)}{f(B)}$$

Illustration with discrete probabilities

- Consider testing for a disease such as covid.
- If D have disease, N do not have the disease, with probabilities $P(D)$ and $P(N)$ where $P(D) + P(N) = 1$.
- If TP test shows positive (suggests you have the disease), TN the test shows negative. $P(TP) + P(TN) = 1$.
- The medical literature reports the sensitivity, $P(TP \mid D)$, and specificity, $P(TN \mid N)$, of the test, which are easy to measure. Take a sample that you know have the disease and see what $\%TP$. Similarly with a sample that you are sure do not have the disease, see what $\%$ TN.
- Diagnosis requires the predictive values, $P(D \mid TP)$ and, $P(N \mid TN)$.
- You can relate specificity and sensitivity to predictive value if you know the prevalence of the disease, the proportion of the population with it, $P(D)$. But this can be difficult to estimate if many who have it show no symptoms.

Calculating predictive values with Bayes

- Suppose we know prevalence is 1%, $P(D) = 0.01$. We have a test that is 99% accurate: $P(TP \mid D) = P(TN \mid N) = 0.99$.
- What is $P(D \mid TP)$? If you test positive, what is the probability you have the disease?
- Bayes Theorem gives this

$$P(D \mid TP) = \frac{P(TP \mid D)P(D)}{P(TP)}$$

- Where

$$\begin{aligned} P(TP) &= P(TP \mid D)P(D) + P(TP \mid N)P(N) \\ &= 0.99 \times 0.01 + 0.01 \times 0.99 = 2 \times 0.99 \times 0.01 \end{aligned}$$

- So

$$P(D \mid TP) = \frac{P(TP \mid D)P(D)}{P(TP)} = \frac{0.99 \times 0.01}{2 \times 0.99 \times 0.01} = 0.5$$

Calculating predictive values numerically

- It is often clearer to use numbers than probabilities.
- Take a population of 100,000. $(1000 = 0.01 \times 100,000)$ have the disease and 99,000 do not.
- Of those with it, $(990 = 0.99 \times 1000)$ test positive, 10 test negative.
- Of those without it $(990 = 0.01 \times 99,000)$ also test positive, 98010 test negative.
- Half of the 1980 who tested positive have the disease.
- 10 out of 98,020, who test negative have the disease.
- We could represent the joint and marginal frequencies as a table.

	D	N	
TP	990	990	1,980
TN	10	98,010	98,020
	1,000	99,000	100,000

Bayes rule

- Apply Bayes theorem to the parameter θ and the data Y.

$$f(\theta \mid Y) = \frac{f(Y \mid \theta) f(\theta)}{f(Y)}$$

- For data Y and random parameter θ, Bayesians derive the posterior distribution, $f(\theta \mid Y)$, as proportional to the product of the likelihood, $f(Y \mid \theta)$ and the prior distribution, $f(\theta)$, treating $f(Y)$ as a constant (it has no information about θ, so can be ignored).

$$f(\theta \mid Y) \propto f(Y \mid \theta) f(\theta).$$

- We will usually write the $\propto$ as $=$.
- Need priors, $f(\theta)$. Can use uninformative priors but they are likely to be improper: $f(\theta)$ does not integrate to one.
- Conjugate priors are widely used because when combined with the likelihood they give a posterior with the same form of distribution.

Posterior distribution

- Bayes rule gives us a posterior distribution for the parameter conditional on the data. Choice of an estimator is a decision problem.
- To choose an estimator we need a loss function. Quadratic loss function gives mean, absolute loss function gives median.
- Suppose that we choose the mean, this is

$$E(\theta) = \int \theta f(\theta \mid Y) d(\theta)$$

  which involves integration, over the support of θ. Similarly, estimating the posterior variance to get a standard error involves integration.
- These integrals can rarely be worked out analytically, instead done numerically through Markov Chain Monte Carlo, MCMC, methods. Increased computing power has made Bayesian methods easier to apply.

Regression

- We can get analytical results for the normal LRM

$$y \sim N(X\beta, \sigma^2 I) = f(y \mid X\beta, \sigma^2 I)$$

- For X independent of β, σ^2 can condition on it, so Bayes rule is

$$f(\beta, \sigma^2 \mid y, X) = \frac{f(y \mid X\beta, \sigma^2 I) f(\beta, \sigma^2)}{f(y \mid X)}$$

- Often more convenient to work with the precision, the inverse of the variance, $h = \sigma^{-2}$ or cov matrix: $H = V^{-1}$.
- Normal gamma prior

$$f(\beta, h) = N(\beta \mid \underline{\beta}, \underline{H}) f_\gamma(h \mid \underline{\sigma^2}, \underline{\nu}_\sigma)$$

with for β prior mean $\underline{\beta}$, prior precision $\underline{H}$ and for σ^2 $\underline{\sigma^2}, \underline{\nu}_\sigma$.

- Together with the normal likelihood this gives a normal gamma posterior

Estimation

- For regression model $y = X\beta + u$ with least squares estimates $\hat{\beta} = (X'X)^{-1}X'y$, $H(\hat{\beta}) = \hat{\sigma}^{-2}(X'X)$ prior mean $\underline{\beta}$, prior precision $\underline{H}$
- The posterior is normally distributed with a mean which is a matrix weighted average

$$\overline{\beta} = (H(\hat{\beta}) + \underline{H})^{-1}(H(\hat{\beta})\hat{\beta} + \underline{H}\underline{\beta})$$

  and precision $\overline{H} = (H(\hat{\beta}) + \underline{H})$.
- Elements of a matrix weighted average vector, do not have to lie between the prior and OLS, $\overline{\beta_i}$ need not be between $\underline{\beta}_i$ and $\hat{\beta}_i$.As $T \longrightarrow \infty$ $H(\hat{\beta})$ gets larger (variance gets smaller), while $\underline{H}$ is constant so asymptotically $\overline{\beta}$ goes to the Maximum Likelihood estimator $\hat{\beta}$.

Shrinkage estimators

- This form of the Bayesian regression estimator

$$\overline{\beta} = (H(\hat{\beta}) + \underline{H})^{-1}(H(\hat{\beta})\hat{\beta} + \underline{H}\underline{\beta})$$

  can be interpreted as a shrinkage estimator (like Ridge Regression or Lasso)
- It shrinks the least squares estimator to the prior $\underline{\beta}$, which could be zero.
- It is like combining estimates of β from two different samples.
- You can implement Bayesian estimation by generating data from the prior (e.g. a DSGE model) and adding it to the real data.

Bayesian VARs

- VARs are just linear regressions so the results for Bayesian regression can be applied. The problem with VARs (Very Awful Regressions, Zellner) is that they have too many parameters, this means that they do not forecast well.
- Sensible, simple, parsimonious models, with few parameters tend to forecast better.
- The very popular Minnesota (Litterman) Priors, treat all the variables as random walks. Prior: coefficient of the lagged dependent variable is one all other coefficients zero.
- The tightness of these priors can be varied. In EViews Mu1 is the prior for the AR1 coefficient, lambda1 the tightness of the prior for the AR1 coefficient, the smaller lambda1 the tighter the prior, zero imposes the prior,.inf gets you an uniformative prior, lambda2 the tightness of the other variables, lambda3 the tightness of the lags

Next time

- The three main causes of the failure of the exogeneity assumption are
  - Simultaneity, two way causation
  - Omitted variables correlated with the included variables
  - Errors in variables
- Next time look at measurement errors, including when we have unobserved "rational" expectations of the future