Connecting Bayesian and Frequentist reporting; Regression

Christopher A. Sims
Princeton University
sims@princeton.edu

October 2, 2020

Likelihood asymptotics for frequentists

- BvM type results imply that in large samples, with a correct model and smooth likelihood, Bayesian posteriors are Gaussian in shape, centered close to the true value of the parameter.

- Asymptotic analysis of pre-sample probabilities implies that the MLE $\hat{\beta}$ satisfies

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \Omega) \,, \tag{1}$$

where, as in the BvM result,

$$\Omega = -\left(E\left[\frac{\partial^2 \log(LH(\beta_0))}{\partial\beta\partial\beta'} \right] \right)^{-1} \,. \tag{2}$$

Bayesian interpretation of frequentist reports

- Likelihood based frequentist estimators and confidence sets will be similar to Bayesian posterior means and credible sets in well-behaved large samples, assuming both types of inference use the same model.

- Note, though, that in our discussion of the CLT for a sample mean, we needed to assume only finite variance, not that X_i were normally distributed, while Bayesian inference leads to the sample mean as an estimator only under a normality assumption.

Conditioning on estimators

- Suppose a Bayesian and a frequentist agree that the data are probably not exactly normal, and the frequentist presents the Bayesian with a sample mean and sample standard error for the mean, estimating the parameter both econometricians are interested in.

- Because the Bayesian agrees that in large samples the CLT may well be a good approximation, she can base approximate post-sample inference on the value of the estimator itself.

- The Bayesian, who would use a different estimator if the underlying data were available to her, can nonetheless treat $\beta \mid \hat{\beta}$ as $N(\hat{\beta}, \Omega)$. (Proving this requires careful specification of regularity conditions.)

Sandwich inference

- The MLE $\hat{\beta}$ under the usual regularity conditions is asymptotically equivalent to the Bayesian posterior mean. That is, $\sqrt{N}$ times their difference converges in probability to zero.

- The MLE satisfies

$$\sum_{i=1}^{N} \frac{\partial \log p(Y_i, \beta)}{\partial \beta} = 0\,. \tag{3}$$

- $$E\left[\frac{\partial \log p(Y_i, \beta)}{\partial \beta} \mid \beta\right] = 0\,. \tag{4}$$

This follows from

$$E\left[\frac{\partial \log p(Y_i,\beta)}{\partial \beta}\right] = \int \frac{1}{p(Y_i,\beta)}\frac{\partial p(Y_i \beta)}{\partial \beta} p(Y_i,\beta)\,dY_i = \frac{\partial}{\partial \beta}\int p(Y_i,\beta)\,dY_i. \tag{5}$$

The derivative of the log likelihood with respect to β is called the **score**.

- Taking a Taylor expansion of the derivative of the log likelihood around $\beta = \beta_0$,

$$0 = \sum_{i=1}^{N} \frac{\partial \log(p(Y_i,\beta_0))}{\partial \beta} + \sum_{i=1}^{N} \frac{\partial^2 \log p(Y_i,\beta)}{\partial \beta \partial \beta'}(\hat{\beta} - \beta_0) + \text{ remainder.} \tag{6}$$

- Assuming the accuracy of the Taylor approximation (which skips the hard part of the argument), we arrive at

$$\sqrt{N}(\hat{\beta} - \beta_0) \doteq - \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \log p(Y_i, \beta_0)}{\partial \beta \partial \beta'} \right)^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \log(p(Y_i, \beta_0))}{\partial \beta} . \tag{7}$$

- The first of the two terms on the right-hand side is the inverse of the time average of a matrix (we assume) has finite, non-singular mean. By the SLLN the time average converges in probability to its expectation, and the matrix inverse is a continuous function of the matrix itself (at points of non-singularity). We have been calling this first term Ω.

- The second term is the sum of i.i.d., mean-zero random variables that (we assume) have a finite variance matrix, which we call B.

- Now we use a fact we haven't previously noted: If $X_i \xrightarrow{P} K$, a constant, and $Y_i =\xrightarrow{D} Z$, then $XY \xrightarrow{D} KZ$. Applying this, we see that

$$\sqrt{N}(\beta - \beta_0) \xrightarrow{D} N(0, \Omega B \Omega)\,. \tag{8}$$

- If the model is correct, it can be shown that $B = \Omega^{-1}$. But both B and Ω can be estimated from the data, and the $\Omega B \Omega$ form, called "the sandwich estimator" for the covariance matrix, is asymptotically justified even if the model is incorrect.

Pros and cons of the sandwich

- Ω is a function of β. Estimating it on the assumption of the model being correct adds no new parameters.

- To estimate B and Ω without assuming model correctness, we generally are adding many new parameters to be estimated.

- Thus even though the sandwich is *asymptotically* more accurate in the presence of model misspecification, it is in finite samples less precisely, perhaps much less precisely, estimated than $\Omega(\beta)$.

- Use of the sandwich should be a judgment call, depending on how big the sample size is and how badly mis-specified we think the model might be. But in fact the sandwich is often simply used as a default.

A rule of thumb

- If $\Omega B \Omega$ and Ω are close, the model is probably not badly misspecified, which suggests using Ω.

- If they are far apart, that could be an indication of serious mis-specfication (or of serious inaccuracy in the sandwich for this sample size). If the model is badly mis-specified, is the MLE or posterior mean of β based on the model still of any use? Is there a way to expand the model — allow for different distributions, nonlinearity, etc., to make it more accurate? If so, this is better than simply reporting the sandwich.

Linear Regression

- We'll look at two approaches to justifying and interpreting linear regression.

- Model-based: Start with a probability model with unknown parameters, show that least squares estimation is optimal, and derive uncertainty measures from the model.

- Design-based: Start with the OLS estimator, ask what it is estimating and what minimal assumptions we need to make what it estimates interesting. Generate distribution theory for the estimator, not the parameter.

Model based

$$y_i \mid \left\{X_j, j = 1, \ldots, N, \beta, \sigma^2\right\} \sim N(X_i\beta, \sigma^2)\,. \tag{9}$$

We'll consider first the case of observations independent across i, conditional on all the X_j's. The X_i's are vectors of length k. For this approach, since we are conditioning on all the values of X_i in the sample, we do not need any assumptions about the distribution of X_i.

Likelihood function:

$$(2\pi)^{-Nk/2}\sigma^{-N}\exp\Big(-\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - X_i\beta)^2\Big)\,. \tag{10}$$

Posterior

Let $\underset{N\times k}{X}$ be the matrix with typical row X_i and y be the correspondingly "stacked" vector of y_i values. Then with a flat prior on β and a flat prior on $\log\sigma^2$ the posterior for β is Normal-inverse-gamma. That is,

$$\beta \mid \{y, X, \sigma^2\} \sim N(\hat{\beta}, \sigma^2(X'X)^{-1}) \tag{11}$$

$$\sigma^2 \mid \{y, X\} \sim \text{Inverse-Gamma}\Big(\frac{n-k}{2}, \frac{s^2}{2}\Big), \tag{12}$$

where $\hat{\beta} = (X'X)^{-1}X'y$ is the ordinary least squares (OLS) estimator of β and $s^2 = (y - X\beta)'(y - X\beta)/(N-k)$. These results can be derived by the same sort of "completing the square" argument that we used in deriving the posterior for a normal mean. To make a draw from this posterior, one draws a σ^2 value as the inverse of a Gamma draw, then conditions on that to make a normal draw from (11).

Design-based

- Suppose we think of our sample $\{y_i, X_i, i = 1, \ldots, N\}$ as a sample from an underlying population, which might be finite or might be continuously distributed.

- Suppose we want to predict y_{N+1} from X_{N+1}, with squared-error loss, in a new draw from the same population, and we are restricted to predictors linear in X_{N+1}.

The estimator

- This means we want to minimize over β, using the joint distribution, (i.e. *not* conditioning on X)

$$E[(y_j - X_j\beta)^2] = E[y_j^2] - 2E[y_jX_j\beta] + E[\beta'X_j'X_k\beta]\,. \tag{13}$$

- Applying calculus we arrive at the optimal

$$\beta^* = \big(E[X_i'X_i]\big)^{-1}E[X_i'y_i]\,. \tag{14}$$

Of course we don't know the expectations in this formula, but we can use the SLLN to take sample averages and estimate them consistently as $X'X/N$ and $X'y/N$. Plugging these into the formula for β^*, the N's cancel and we are back to $\hat\beta = (X'X)^{-1}X'y$, the OLS estimator.

Asymptotic distribution for the estimator

- Let $\varepsilon_i = y_i - X\beta^*$. Then

$$\hat{\beta} = (X'X)^{-1}X'(X\beta^* + \varepsilon_i) = \beta^* + (X'X)^{-1}X'\varepsilon\,.. \qquad (15)$$

- Our assumptions and definition of β^* guarantee that $E[X'\varepsilon] = 0$. This does not guarantee that $\hat{\beta}$ is unbiased, though, because our assumptions do not rule out dependence between $X'X$ and $X'\varepsilon$.

- Nonetheless the CLT, under the assumption that $X_i'\varepsilon_i$ has a finite covariance matrix, guarantees that

$$\frac{1}{\sqrt{N}}X'\varepsilon \xrightarrow{D} N(0, \mathrm{Var}(X'\varepsilon))\,. \qquad (16)$$

Design-based asymptotics continued

We can consistently estimate the limiting variance:

$$\frac{1}{N}\sum_{i=1}^{N} X_i' X_i \hat{\varepsilon}_i^2 \xrightarrow{P} \mathrm{Var}(X_i' \varepsilon_i) \ , \tag{17}$$

where $\hat{\varepsilon}_i = y_i - X_i \hat{\beta}$. This depends on the terms in the sum having finite expectation, which can be restrictive in applications. If the X_i's have a fat-tailed distribution, as might be expected with income or firm-size data for example, the convergence does not hold.

What is done in practice

We end up using the “heteroskedasticity-robust” version of a sandwich estimator: estimate β by OLS, use

$$(X'X)^{-1} \sum \Big(X_i' X_i \hat{\varepsilon}_i^2 \Big) (X'X)^{-1} \tag{18}$$

as the covariance matrix of $\hat{\beta}$. Note that if $E[\varepsilon_i^2 \mid X_i] = \sigma^2$, a constant, $\mathrm{Var}(X_i'\varepsilon_i) = \sigma^2 E[X_i' X_i]$, which brings us back to $\hat{\sigma}^2 (X'X)_{-1}$ as the variance matrix estimate. ($\hat{\sigma}^2$ is the sample standard deviation of ε_i.)

Which is better?

- The model, if its assumptions are satisfied, delivers a more useful result: It implies that for an arbitrary X_j, not necessarily observed in the sample, $X_j\beta$ is the "best" predictor of y_j, where "best" means minimizing expected RMSE conditional on the data.

- Its results do not depend on the sample being "large", and they are directly post-sample distributions.

- The design-based theory avoids the assumption that $E[y_i \mid X_i] = X_i\beta]$, but if this does not hold, why are we interested in β?

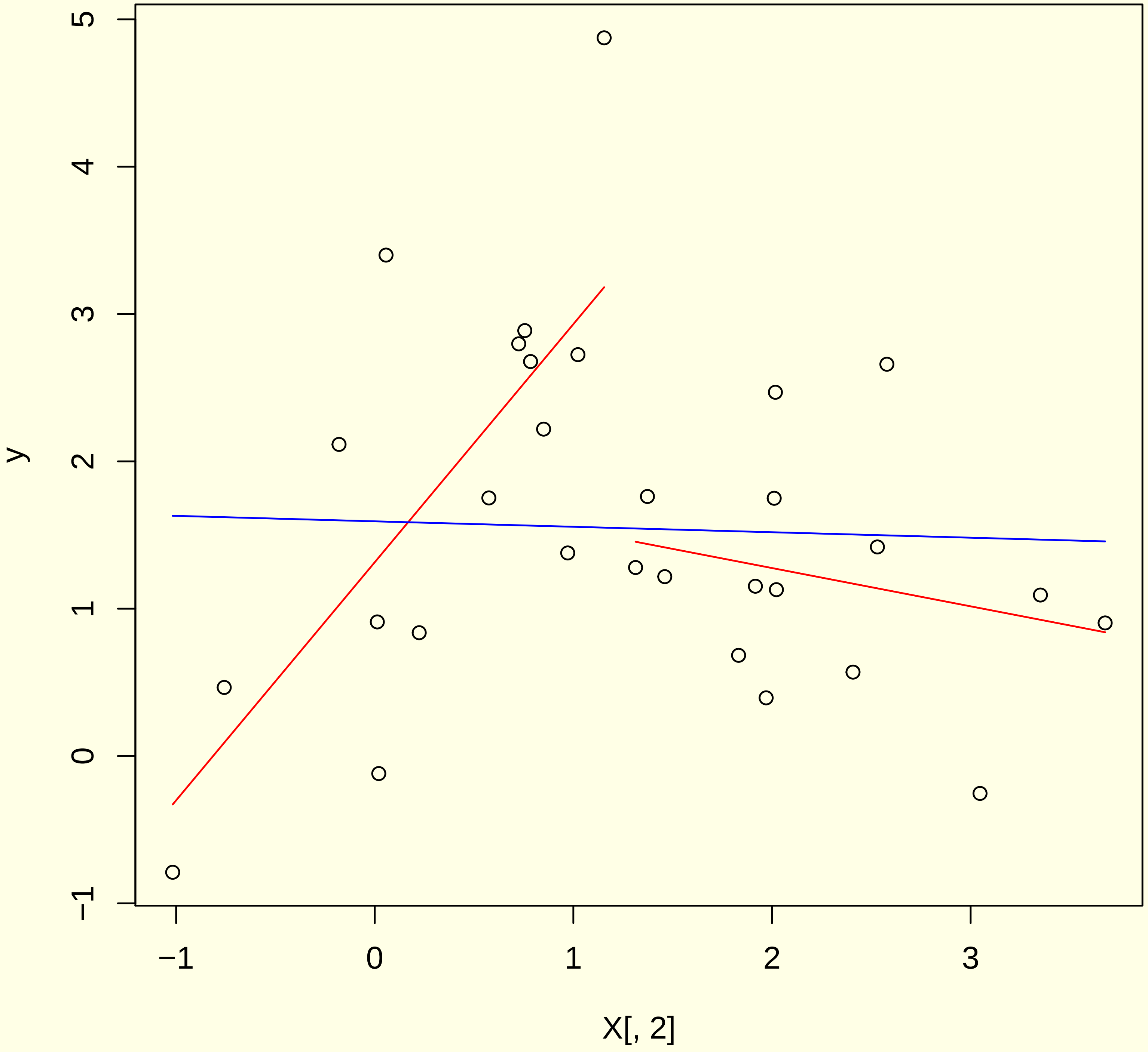
- The design-based distributions are all pre-sample, though as usual we may guess that our sample is “large” so that we can invert the asymptotic distribution theory to obtain approximate post-sample probabilities.

Prediction with design-based assumptions

- The assumptions justifying design-based theory imply that the linear model we are estimating provides the best RMSE prediction of y_i *within the class of linear predictors*, and *assuming that in our predictions we will draw repeatedly from the same distribution of X_i values that generated our sample*.

- Because $E[y_i \mid X_i]$ may be nonlinear in X_i under these assumptions, if we are making predictions with X_i's that are mostly larger than those in the sample, the predictions may be much worse than would be possible even with a linear model.

Fitting a nonlinear E[y|X]



Improving the standard normal linear model

- The SNLM assumes $E[y_i \mid X_i]$ is linear, that the distribution of the residuals $\varepsilon_i \mid X_i$ is normal, and that $\sigma^2 = E[\varepsilon_i^2]$ is constant.

- These assumptions are all often violated in in applications, and the design-based distribution theory does not depend on them.

- But we can relax these assumptions and still have an estimable, and more useful, model.

Nonlinearity

- Use powers of X_i as regressors (polyomial fit).

- Use piece-wise linear model, or more general splines.

- These add parameters, so we have to think about where to stop; if we add enough parameters the frequentist distribution theory falls apart, and specifying reasonable priors in high-dimensional parameter spaces is hard.

Heteroskedasticity

- We can model the form of dependence of $E[\varepsilon_i^2 \mid X_i]$ on X. For example, sometimes data are results of surveys on units of different size. Then the sampling error might be a function of unit size, decreasing in unit size. Or large units might have larger disturbances.

- Fitting $\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 Z + \alpha_1 / Z$, where Z is a measure of unit size is often reasonable.

Generalized least sauares

$$y \mid (X,\beta) \sim N(X\beta, V) \Rightarrow \beta \mid (y, X, V) \sim N\big(\hat{\beta}_{GLS}.(X'V^{-1}X)^{-1}\big) \quad (19)$$

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y \quad (20)$$

Since V depends on unknown parameters, we have to estimate it, and the full posterior p.d.f. for those parameters is not a standard distribution like Inverse-Gamma.

In practice often V is estimated by some consistent method and then conditioned on as if it were known exactly, but with Markov Chain Monte Carlo (MCMC) methods, use of the full, non-standard posterior distribution is not hard.

Non-normality

(in Lecture 11)

Asymptotic improvements?

(in Lecture 11)