Stretching the regression model

Christopher A. Sims
Princeton University
sims@princeton.edu

October 6, 2020

One more point about GLS

- Recall that $V = E[\varepsilon\varepsilon' \mid X]$, and that the GLS formula, for the flat-prior posterior distribution $\beta \mid V$, is

$$\beta \mid \{Y, X, V\} \sim N\Big((X'V^{-1}X)^{-1}X'V^{-1}y, (X'V^{-1}X)^{-1}\Big) \,. \tag{1}$$

- For any positive definite V we can find a matrix W satisfying $W'W = V^{-1}$.

- Let $X^* = WX$ and $y^* = Wy$. Then applying the usual OLS formulas to X^* and y^* delivers the GLS formulas.

- For i.i.d. data, V is diagonal, so W has a simple form: it is diagonal, with the inverses of the residual standard deviations on the diagonal.

- Thus GLS in this case is “weighted least squares”, with the observations with larger residual variances getting less weight.

Frequentist asymptotics when $E[y_i \mid X_i] = 0$

- We've discussed finite-sample inference in the standard normal linear regression model and asymptotic distribution theory for OLS assuming only an i.i.d. sample and existence of $E[X_i' X_i]$ and $E[\varepsilon_i^2 X_i' X_i]$.

- There is an intermediate case for frequentist asymptotics, where we add the assumptions $E[\varepsilon_i \mid X_i] = 0$ and $E[\varepsilon_i^2 \mid X_i] = \sigma^2$.

- Under these assumptions $\hat{\beta}_{OLS}$ is unbiased, and we get asymptotic normality with the usual covariance matrix, without assuming normality.

Details

$$\begin{aligned} E[\hat{\beta}_{OLS} \mid X] &= E[(X'X)^{-1}X'(X\beta + \varepsilon) \mid X] \\ &= \beta + (X'X)^{-1}X'E[\varepsilon \mid X] = 0\,. \end{aligned} \quad (2)$$

This unbiasedness result does not require that $E[X_i'X_i]$ be finite, but it does require the i.i.d. assumption, so that $E[\varepsilon_i \mid X_i] = 0 \Rightarrow E[\varepsilon \mid X] = 0$.

$$\begin{aligned} \sqrt{N}(\hat{\beta}_{OLS} - \beta) &= \left(\frac{1}{N}X'X\right)^{-1}\frac{1}{\sqrt{N}}X'\varepsilon_i \\ &\xrightarrow{D} (E[X_i'X_i])^{-1}N(0, \sigma^2 E[X_i'X_i]) = N(0, \sigma^2(E[X_i'X_i])^{-1}). \end{aligned} \quad (3)$$

Non-normality

- Last lecture we discussed expanding the linear regression model by allowing for non-lineearity (as you are doing in this week's exercise) and allowing for heteroskedasticity via GLS.

- One can allow for nonlinearity via adding nonlinear functions of X_i to the regression, and one can allow for heteroskedasticity by estimating a function $\sigma^2(X_i, \alpha)$ and using it to generate a V matrix for the GLS formula.

- If nonlinearity is present, and we fail to allow for it, we fail to get valid estimates of $E[y_i \mid X_i]$

- If heteroskedasticity is present and we fail to allow for it, we get incorrect estimates for $\mathrm{Var}(\beta \mid y, X)$ or $\mathrm{Var}(\hat{\beta} \mid \beta, \sigma^2)$.

Non-normality may be less important

- The frequentist theory on the previous slide shows that if non-normality is present, even if we fail to allow for it, we get approximately valid inference in large samples.

- This “approximate validity” is, from a Bayesian perspective, only for inference conditional on $\hat{\beta}_{OLS}, s^2$, rather than conditional on the full sample, if non-normality is present.

- If the sample size is modest, or the residuals are strongly non-normal, or we have good a priori reason to think a certain non-normal distribution applies, allowing for the non-normality may be important.

- Correctly modeling non-normality of residuals can deliver improved estimates, but incorrectly modeling it can undermines the asymptotic normality result.

- If you want to understand this at a deeper level, look for literature on "semi-parametric efficency bounds", particularly classic papers by Gary Chamberlain.

Accounting for outliers

- Quite often, especially with fine-time unit or financial data, it is clear that there are too many large “outlier” residuals to be consistent with a normality assumption. (E.g., a normal q-q plot gets steep at both ends).

- One possibility that should never be ignored is that the outliers are some sort of data error. If the errors can be tracked down, the data can be repaired or the damaged observation removed.

t or double-exponential errors

- But if data errors are not the problem, modeling the residuals as t distributed or double-exponential distributed may improve the fit.

- Both the t likelihood and the double-exponential likelihood lead to a version of “weighted least squares” in which large-absolute-value residuals are downweighted.

- The double exponential leads to Least Absolute Error (LAE) estimation for the MLE, which can be computed efficiently via linear programming. However the posterior density is messy and only asymptotic distribution theory is available for a frequentist approach.

- The t posterior can be computed efficiently via Markov Chain Monte Carlo (MCMC) methods, which we may have time to discuss in Lecture 12.

Model comparison

- We're considering a mixed continuous-discrete distribution for parameter values.

- In particular, we're considering the case where there is one parameter m, called model number that takes on M discrete values, and other parameters, collected in a vector β, that are continuously distributed.

- We would like to start with some prior beliefs about m, say equal probabilities on its M possible values, then observe some data Y and update our beliefs about m to a distribution for $m \mid Y$.

Forming the posterior on m

- We'll assume there is a marginal prior $\mu(m)$ over models and a conditional prior $\pi(\beta \mid m)$ over β for each model. The model is $p(Y \mid \beta, m)$. Then the posterior over m and β is proportional to

$$\mu(m)\pi(\beta)p(Y \mid \beta, m)\,. \tag{4}$$

- This implies that the marginal posterior for m is proportional to

$$\mu(m)\int \pi(\beta \mid m)p(Y \mid \beta, m)\,d\beta\,. \tag{5}$$

The marginal data density

- Usually there is a separate β_m parameter vector for each model and priors for β_m and β_q are independent for $m \neq q$. Parameters not in β_m then enter (5) only via their prior densities, and integrate out of the expression, so it becomes

$$\mu(m) \int \pi(\beta_m \mid m) p(Y \mid \beta_m, m) \, d\beta_m \, . \tag{6}$$

- This object is known as the **marginal data density** or mdd.

- In normal linear regression models with conjugate priors, or where we are approximating the posterior within each model with a normal density, there are analytic expressions for it, but often it requires numerical integration.

From mdd to BIC

Suppose we are satisfied with the asymptotic Gaussian approximation to the shape of the posterior density. In that case, we can calculate mdd's analytically, using just the log posterior density at its maximum $\hat{\beta}$ and V, the second derivative matrix for the log posterior density at that point. Let k be the number of parameters in the model. We temporarily suppress the m subscripts as we're dealing with a single model, and use the notation $\psi(Y, \beta) = \mu(m)\pi(\beta_m \mid m)p(Y \mid \beta_m)$. The mdd integral (6) then evaluates to

$$\log\big(\psi(Y \mid \hat{\beta})\big) + \frac{k}{2}\log(2\pi) - \frac{1}{2}\log(\big(|-V|\big)\,. \tag{7}$$

Log odds ratios across models, and from them the whole posterior distribution across models, are given by the differences across models of (7).

Shortcuts

- Equation (7) can be evaluated fairly easily. Usually $\hat{\beta}$ and $\log p(Y \mid \beta)$ are calculated as the model is estimated. Calculating V for a large model can be challenging.
- Notice, though, that

$$V = \frac{\partial^2 \log p(Y \mid \hat{\beta})}{\partial \beta \partial \beta'} + \frac{\partial^2 \log\big(\mu(m)\pi(\beta \mid m)\big)}{\partial \beta \partial \beta'} \,. \tag{8}$$

The second term in this expression does not change with sample size. The first term is the average over the sample of individual observation second derivatives, and thus as we have observed before converges in probability, when divided by N, to what we have called $-\Omega^{-1}$.

BIC

- Suppose that in (7) we replace $-V$ with $N\Omega^{-1}$ and drop all terms in the resulting expression that don't change with N.

- This gives us
$$\log\bigl(p(Y \mid \hat{\beta})\bigr) - \frac{k}{2}\log N\ . \tag{9}$$
The Ω has disappeared, because its term does not change with N.

- So now to compare models with BIC, we take differences of these expressions across models. Assuming the usual regularity conditions, BIC is guaranteed eventually to be largest for the true model.

Caveats

- While this is a handy shortcut, it is quite rough. When k is large, the $k\log(2\pi)$ terms, and especially the $\log|\Omega|$ terms that are dropped can be large.

- And from (5), we see that no matter what the sample size, if we change $\mu(m)$, we change mdd's proportionately — the effect of the prior over models does not die away with N. μ drop out of BIC, because BIC is not an approximation of posterior log odds ratios: it is just a quantity that eventually, in large enough samples, favors the true model by an arbitrarily large margin.

Comparison to standard F and χ^2 tests

- In a standard normal linear regression model, the BIC criterion, applied to a linear restriction on the model, is equivalent to comparing the F statistic for the restriction to $\log N$.

- Except possibly at very small sample sizes, this favors H0 (the restricted model) much more strongly than the F test at standard 1% or 5% levels.

- The critical value of the F statistic at a given level gets smaller, the more extra degrees of freedom are in the larger model, while BIC does not.

Choosing among many models

- Several popular machine learning algorithms (random forests, LASSO, e.g.) aim at choosing a model with a modest number of explanatory variables when a very large number are a available.

- Posterior odds provide an approach to this task. Posterior odds generally require more computation than the ML algorithms, but they provide more useful output.

- Two interesting papers applying these methods to many-regressors situations in economics are Fernández et al. (2001) and Giannone et al. (2020).

*

References

FERNÁNDEZ, C., E. LEY, AND M. F. J. STEEL (2001): "Model Uncertainty In Cross-Country Growth Regressions," *Journal of Applied Econometrics*, 16, 563–576.

GIANNONE, D., M. LENZA, AND G. E. PRIMICERI (2020): "Economic Predictions With Big Data: The Illusion of Sparsity," Tech. rep., Northwestern University.