EVALUATING COVID QUANTILE FORECASTS: AN R VINAIGRETTE

ROGER KOENKER

ABSTRACT. Some notes on evaluation of quantile forecasts for Covid-19 and related uncertainties.

1. Introduction

Motivated by some email conversations with Ryan Tibshirani last summer, I've been curious about how Nick Reich's Forecast Hub `https://covid19forecasthub.org` creates their ensemble forecasts and how they are evaluated. I'll attempt to describe evaluation first and then try to deal with ensembles.

2. Scoring Rules for Quantile Forecasts

There is quite an extensive emerging literature on forecast evaluation heavily influenced by Gneiting and Raftery (2007), which tries to come to grips with something more than point forecasts and the dreaded mean squared error. Density forecasts are now commonplace in econometrics, epidemiology and other fields, so one might well ask: Given a forecast density, f, and a single observation, y, how should I evaluate the forecast? In Gneiting-Raftery terminology a strictly proper scoring rule for for a density forecast is one that would induce the forecaster to report the true density presuming that it were known. Thus, for example, the logarithmic score,

$$\log S(f,y) = \log f(y)$$

is strictly proper since it is minimized when the true density is g, by choosing $f = g$. Unless one would like to restrict attention to parametric models, such scoring rules impose a rather heavy burden on the forecaster since they require that f be specified in its entirety. A much simpler strictly proper scoring rule is the quantile rule,

$$S_\tau(f,y) = \rho_\tau(y - Q_f(\tau))$$

where $Q_f(\tau)$ is τth quantile of the f distribution and ρ_τ is the usual check function. This is obviously minimized by choosing f so it has "right" τth quantile. Of course one quantile doesn't provide a very full description of the whole distribution, so it is natural to consider combining several such scoring rules. In particular, we can construct an interval scoring rule,

$$IS_\alpha(f,y) = \rho_\alpha(y - Q_f(\alpha)) + \rho_{1-\alpha}(y - Q_f(1-\alpha))$$

as depicted in Figure 1 for $\alpha = 0.1$ and interval limits $[-1,1]$. For $y \in [-1,1]$ the score is 0.1, and then increases linearly for y outside this interval. Similar scoring functions can be built by concatenating several interval scoring rules.

December 23, 2021. A genre manifesto for R Vinaigrettes is available at `http://davoidofmeaning.blogspot.com/2016/12/r-vinaigrettes.html`.

The Forecast Hub protocol is that those submitting forecasts for consideration should submit a point estimate of the median, μ, and $K = 11$ prediction intervals for each region and forecast horizon with nominal coverages, $1-\alpha_k$, and $\alpha_1 = 0.02, \alpha_2 = 0.05, \alpha_3 = 0.10, \ldots, \alpha_K = 0.90$. An aggregated score is then contructed as described Bracher et al. (2021),

$$WIS_\alpha(f, y) = \left[w_0 |y - \mu| + \sum_{k=1}^{K} w_k IS_{\alpha_k}(f, y) \right] / (K + 1/2)$$

with suggested weights, $w_0 = 1/2$ and $w_k = \alpha_k/2$ for $k = 1, \ldots, K$. This weighted interval scoring rule can be viewed as a measure of distance between the predictive distribution and distribution of the (random variable) y. The red curve superimposed in the figure illustrated one such rule based on five distinct intervals. It approximates the "continuous ranked probability score,"

$$CRPS(F, y) = \int (F(x) - \mathbb{1}(x \geq y))^2 dx,$$

and can be conveniently decomposed to investigate aspects of the forecast distribution that are unsatisfactory. Bracher et al. (2021) suggest several graphical devices for this purpose.

3. Ensemble Forecasting

Over the course of the pandemic the Covid-19 Forecast Hub team has received forecasts in the format specified above from groups of researchers around the globe. In the last week of November, 2021, for example, 50 forecasts were submitted. These can be individually evaluated, but a more challenging and potentially rewarding task is to combine these forcasts into an ensemble forecast that reflects some sort of concensus forecast that can then be promulgated by the CDC. Since forecasts are reported for cases, hospitalizations and deaths at the county, state and national level at one and two week horizons. This represents an enormous data management problem even before any decisions are taken on designing an aggregation mechanism.

Various forms of model selection, model averaging and related ensemble forecasting methods have proliferated, but it seems fair to say that there is no agreed upon general strategy. Over the course of the pandemic several methods have been employed:

- From April 13 to July 21 2020, the ensemble simply computed sample means of the predicted quantiles for each quantile for all eligible models at each location,
- From July 28, sample means were replaced by sample medians.
- From November 15 2021 the ensemble was based on the ten best model submissions in the prior 12 weeks evaluated by their weighted interval score (WIS), with models with better WIS over this period receiving more substantial weight.

See `https://covid19forecasthub.org/doc/ensemble/` for further details.

One can easily imagine other emsemble aggregation methods, but at the scale of the covid forecasting effort it is obviously necessary to exercise considerable restraint before leaping into new procedures. One can view the weighted WIS procedure currently in use as a compromise between the earlier sample mean and sample median procedures. I was motivated to look into this again by the paper of Yao et al. (2018) that reviews the Wolpert (1992) and Breiman (1996) notions of "stacking" from a Bayesian perspective. Before delving into modeling entire distributions, I will try to briefly describe my understanding of how stacking works in conventional regression, point-forecasting applications.

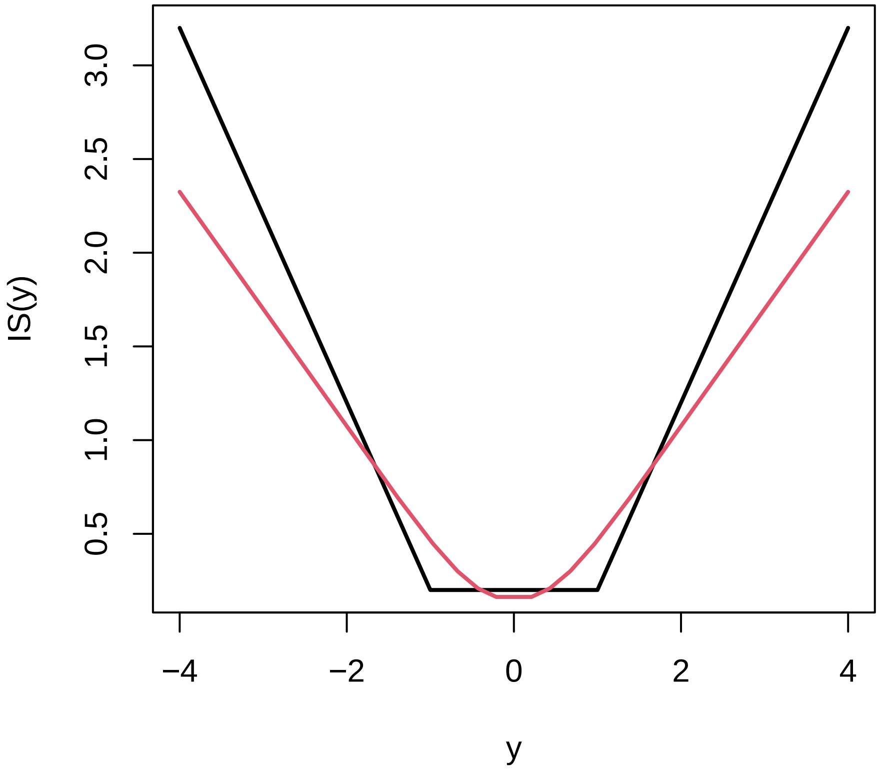

Figure 1. Interval Score Function

3.1. **Stacking for ordinary regression.** Consider the standard regression setting in which we see data consisting of n pairs, (y_i, x_i). We have K candidate models $f_k : k = 1, \ldots, K$. For each model we can compute estimates $\hat{f}_k^{(-i)}(x_i)$ as the leave-one-out prediction of the ith observation from the kth model when we omit the ith observation. Now we solve, for weights $\omega \in \Omega \subset \mathbb{R}^K$,

$$\hat{\omega} = \text{argmin} \sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} \omega_k \hat{f}_k^{(-i)}(x_i))^2,$$

In the Breiman formulation Ω is K simplex, and the positivity constraint helps to ensure that there is a unique solution, and that poor models receive weight zero. Positivity constraints like this also encourage a form of sparsity by this elimination of poor models in the spirit of Breiman's non-negative garotte. Stacking typically performs better from a predictive vantage point than other forms of model averaging. However, as Yao et al remark in its original point estimation form it is not well suited to Bayesian settings where the entire posterior is targeted. To rectify this oversight, Yao et al consider several scoring rules sanctioned by Gneiting and Raftery (2007), but curiously not the WIS rule. I attribute this to the fact that being Bayesian they are addicted to parametric models. Instead they focus on the logarithmic scoring rule that keeps things in the KL playground. This is probably convenient as long as the models under consideration are all parametric, but my intention is to play in the quantile regression sandbox, for which the WIS rules seem opportune.

3.2. **Stacking for Quantile Models.** Leave-one-out fitting for quantile regression isn't enough of a perturbation to accomplish anything useful. Leave-m-out methods might serve better, but are computationally costly. Fortunately, there is no need to evaluate within sample fit for each of the constituent models at the Forecast Hub. This might be prudent for each submitted model *before* submission, but at the Hub there it is only necessary to aggregate based on prior out-of-sample performance of the models, for example by their WIS. Of course there are components being forecast, several time horizons so how this should be done is still problematic. This is all left for another day.

References

Bracher, J., Ray, E. L., Gneiting, T. and Reich, N. G. (2021), 'Evaluating epidemic forecasts in an interval format', *PLOS Computational Biology* .

Breiman, L. (1996), 'Stacked regressions', *Machine Learning* **24**, 49–64.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *J Am Stat Assoc.* **102**, 359–378.

Wolpert, D. H. (1992), 'Stacked generalization', *Neural Networks* **5**, 241–259.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018), 'Using stacking to average Bayesian predictive distributions (with discussion)', *Bayesian Analysis* **13**, 917–1007.