

A comparison of forecasting ability of models for teenage conceptions & MRSA bacteraemia rates

Hayley E Jones
& David J Spiegelhalter

Statistical Laboratory / MRC Biostatistics Unit,
Cambridge, UK

ISCB, Prague.
August 2009

Introduction

- Routine 'performance' data increasingly collected on many healthcare providers at regular intervals.
- Examples:
 - ① Teenage conceptions in English Local Authorities.
 - ② Methicillin Resistant *Staphylococcus Aureus* (MRSA) bloodstream infections in NHS Trusts.
- **Dual objective:** Predictive system that
 - (i) **Provides well-calibrated predictions**, assuming steady state.
 - (ii) Can be used to identify important recent changes, if there are any.

Notation

- Data:

O_{it} = observed cases

E_{it} = 'expected' cases, based on fixed effects regression

- Assume

$$O_{it} | r_{it} \sim \text{Poisson}(r_{it} E_{it})$$

$i = 1, \dots, m$ healthcare providers;

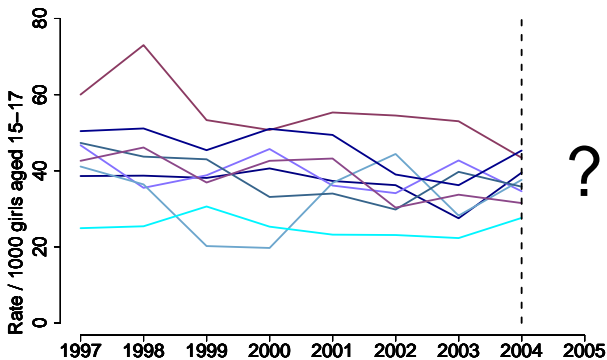
$t = 1, \dots, T$ time periods

- **We reserve the data from the final period, T , to be used for model comparisons.**

Types of smoothing

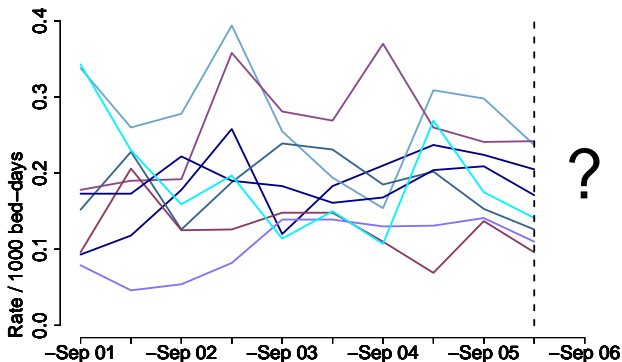
- *Smoothing* observed rates can reduce statistical noise, increase power to detect 'unusual' performance, & correct for 'regression-to-the-mean'.
- But there are several options:
 - 1 **Smooth between units at one time point**
(Hierarchical model)
 - 2 **Smooth within each unit independently over time**
(EWMA / DLM)
 - 3 **Both?**
(‘Bidirectional’ smoothing)

Teenage conceptions data



Data on 8 of 352 Local Authorities.

MRSA data



More volatility, but no evidence of overall change over this time.

Structure of presentation

- 1 Simple example of smoothing between providers at a single time point.
- 2 Simple example of smoothing within each provider independently over time.
- 3 Example of combining the two.
- 4 Tools for comparing predictive ability.
- 5 Results for two data sets & conclusions.

Smoothing between providers at a single time point

e.g. Poisson-gamma model:

$$r_i | \mu, \tau \sim \text{IID Gamma}[\mu, \tau^2].$$

Assuming μ and τ are known, we obtain

Shrinkage estimate:

$$\hat{r}_i = w_i \frac{O_i}{E_i} + (1 - w_i) \mu$$

where

$$w_i = \frac{\tau^2}{\tau^2 + \mu/E_i}$$

- Empirical Bayes: use plug-in estimates of μ and τ .
- Implied predictive distribution for next period is negative binomial.

Smoothing within each provider independently

e.g. EWMA

For simplicity, transform to approx. normality:

$y_{it} \equiv \log(O_{it}/E_{it})$, assume $E(Y_{it}) = \theta_{it}$ & $V(Y_{it}|\theta_{it}) = 1/E_{it}$.

Exponentially weighted moving average (EWMA)

$$\hat{\theta}_{it} = \kappa \hat{\theta}_{i,t-1} + (1 - \kappa) y_{it}$$

for $t = 2, \dots, T - 1$, where $0 \leq \kappa \leq 1$, assuming e.g. $\hat{\theta}_{i1} = y_{i1}$.

- Assumption of underlying model (e.g. normal steady model) needed for full forecasting distributions.
- 'Exact' Poisson versions also available (DGLM / RA-EWMA: Grigg & Spiegelhalter, 2007).

Bidirectional smoothing

In summary:

- 1 Simple hierarchical model fitted to time $T - 1$ data
⇒ prediction for T **shrunk towards average across providers.**
- 2 DLM / DGLM fitted independently to each provider
⇒ EWMA type shrinkage of prediction **towards past observations on that provider.**

What about combining the 2?

Bidirectional smoothing

Hierarchical AR(1) model (Lin *et al.*, 2009)

$$O_{it}|r_{it} \sim \text{Poisson}(r_{it}E_{it})$$

Simple hierarchical model assumed to hold marginally in each time period:

$$\log(r_{it}) \sim \text{Normal}(\mu_t, \tau_t^2) .$$

Time series structure on each standardised process:

$$\frac{\log(r_{it}) - \mu_t}{\tau_t} = \phi \left(\frac{\log(r_{i,t-1}) - \mu_{t-1}}{\tau_{t-1}} \right) + \eta_{it} , \quad t = 2, \dots, T$$

No simple closed form predictive distribution, therefore use MCMC.

Bidirectional smoothing

- Such models, resulting in ‘bidirectional’ smoothing, are increasingly being suggested in the literature (e.g. West & Aguilar, 1998; Van Houwelingen *et al.*).
- But no systematic evaluation has been made, comparing these models to simpler ‘one-way’ smoothing alternatives.
- Tools are needed to **compare the quality of the predictions based on different models.**

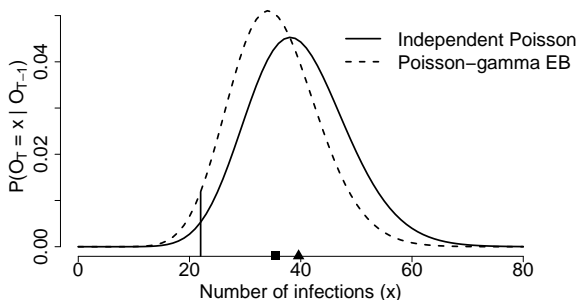
Evaluation criteria

We use 3 approaches, drawing heavily on recommendations of Gneiting *et al.* (2005) in the field of weather forecasting:

- 1 Accuracy of point predictions,
e.g. $MSE = \frac{1}{m} \sum_{i=1}^m (O_{iT} - \hat{O}_{iT})^2$.
But also evaluate the full forecasting *distributions*.
- 2 Uniformity of predictive p -values.
- 3 Proper scoring rules: log-score and CRPS.

Example of 2 predictive densities

Number of MRSA infections in a particular NHS Trust.



▲ Point prediction under independent Poisson model.

■ Point prediction under Poisson-gamma EB model.

Uniformity of predictive p -values

- A correctly calibrated forecasting distribution
≡ Events declared to have probability p occur a proportion p of the time on average.
- If this is the case, then the m predictive p -values should have an approximately $Uniform(0,1)$ distribution.
- Can assess this:
 - 1 **Visually**, using histograms and plots of ordered p -values.
 - 2 **Using test statistics** e.g. Kolmogorov-Smirnov D or Cramér-von-Mises W^2 .

Proper scoring rules

- Uniformity of the predictive p -values is a necessary but not sufficient condition for a forecasting system to be 'ideal' (Gneiting *et al.*, 2007).
- Therefore also consider 2 proper scoring rules.
- 'Proper' \equiv expectation of the score is minimised by the ideal forecasts.

Logarithmic Score

$$LS_i = -\log(f(O_i))$$

- Examine mean of these over providers.

Proper scoring rules

- CRPS has been recommended as a more robust alternative:

Continuous Ranked Probability Score

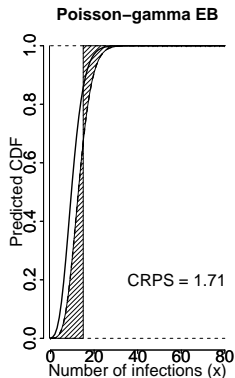
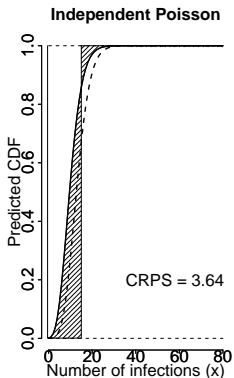
$$\begin{aligned} CRPS(F_i, O_i) &= \int_{-\infty}^{\infty} (F_i(x) - I\{O_i \leq x\})^2 dx \\ &= E|O_i^{pred} - O_i| - \frac{1}{2}E|O_i^{pred} - O_i^{pred'}| \end{aligned}$$

(Gneiting & Raftery, 2007)

- Unlike log score, CRPS is influenced by width of predictive distributions.

Individual CRPS contributions

e.g. Alternative predictive distribution functions for one particular NHS Trust:



Results

Model comparison for the **teenage pregnancies** data:

		<i>MSE</i>	<i>MAE</i>	<i>D</i>	W^2	Width	CRPS	LS
0	Pois indep	249	11.2	0.04	0.08	45.3	8.00	3.98
1	PG EB	197	10.0	0.07	0.41	41.1	7.17	3.87
1	P-LN Bayes	197	10.0	0.06	0.28	40.9	7.16	3.87
2	EWMA	215	10.2	0.08	0.78	43.6	7.20	3.86
2	Pois RW	215	10.2	0.09	1.35	44.1	7.23	3.87
2	Pois AR(1)	201	10.2	0.14	3.44	45.0	7.17	3.87
3	HR AR(1)	189	9.7	0.05	0.33	42.1	6.88	3.83

0) No smoothing, 1) Smooth between, 2) Smooth within, 3) Both.

Results

Model comparison for the **MRSA** data:

		<i>MSE</i>	<i>MAE</i>	<i>D</i>	W^2	Width	CRPS	LS
0	Pois indep	52	5.5	0.09	0.24	18.8	3.81	3.19
1	PG EB	36	4.5	0.06	0.07	17.3	3.15	3.01
1	P-LN Bayes	37	4.6	0.06	0.09	17.4	3.17	3.01
2	EWMA	36	4.6	0.06	0.11	17.0	3.13	3.09
2	Pois RW	37	4.5	0.05	0.11	18.7	3.11	2.99
2	Pois AR(1)	30	4.2	0.08	0.22	18.4	2.86	2.93
3	HR AR(1)	33	4.3	0.06	0.14	18.6	2.96	2.96

0) No smoothing, 1) Smooth between, 2) Smooth within, 3) Both.

Conclusions

- Smoothing observed performance measures has clear benefits in terms of predictive accuracy.
- The 'bidirectional' smoothing model of Lin *et al.* was found to perform particularly well on 2 quite different examples.
- This model is highly interpretable, automatically adapts to characteristics of dataset & is straightforward to program in WinBUGS.
- Seems reasonable to suggest that it should be used as a default.
- However, this model takes a very long time to fit!
- **Future research should focus on faster / simpler methods for fitting bidirectional models.**
e.g. 2-stage approach of Martz *et al.* (1999). Can this perform as well?

References

- O A Grigg & D J Spiegelhalter. A simple risk-adjusted exponentially weighted moving average. *J Am. Statist. Assoc.*, 102:140-152, 2007.
- R Lin, T A Louis, S M Paddock & G Ridgeway. Ranking USRDS provider-specific SMRs from 1998-2001. *Health Services & Outcomes Research Methodology*, 9:22-38, 2009.
- M West & O Aguilar. Studies of quality monitor time series: the V.A. hospital system. *Technical report*. Duke University. 1998.
<http://ftp.isds.duke.edu/WorkingPapers/97-22a.pdf>
- H C van Houwelingen, R Brand & T A Louis. Empirical Bayes methods for monitoring health care quality. *Technical report*.
www.msbi.nl/dnn/People/Houwelingen/Publications/tabid/158/Default.aspx
- T Gneiting, A E Raftery, A H Westveld III & T Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133:1098-1118, 2005.
- H F Martz, R L Parker & D M Rasmuson. Estimation of trends in the scram rate at nuclear power plants. *Technometrics*, 41:352-364, 1999.