Biometrics 0, 1–5 DOI: 0000-0000-0000
January 2024

Extension of People-Like-Me Methods Penalization, Prediction, and Beyond

Randy Jin

Department of Biostatistics and Bioinformatics, University of Colorado Anschutz Medical Campus

*email: xin.2.jin@cuanschutz.edu

and

Elizabeth Juarez-Colunga

Department of Biostatistics and Bioinformatics, University of Colorado Anschutz Medical Campus

*email: elizabeth.juarez-colunga@cuanschutz.edu

Summary: The text of your summary. Should not exceed 225 words.
Key words: inversed-distance weightcurve matchingpredictive interval.

This paper has been submitted for consideration for publication in *Biometrics*

1. 1. Introduction

Your text comes here. Separate text sections with

2. 2. Methods

Text with citations by Heagerty et al. (2000), (Pepe, 2003).

The prediction horizon extends to the

2.1 *2.2 The predictive coverage interval*

The problem for the predictive interval with gamlss model:

Q1. The predictive interval is symmetric, but in fact is a not a requirement for the prediction interval based on our selected matches. The systematic bias introduced by the asymmetric distribution for the matches observations; The systematic bias introduced by the density of the given time points.

(using weights can be a solution for the systematic bias, but we need to be careful about the weights we used.) for example, imagine the tallest person in the study, that all the other people served as matches would be shorter than this target. The prediction based on such matches would be still underestimated. However, we can use the weights to adjust the prediction based on the distance. which means the prediction will be more close too potentially the second tallest individual (the tallest one in the training set). Even though, we cannot accurately predict for the target, but we can extract certain information which provided by what we know, other than nothing. But we still need to find a way to quantify such limitation.

This means for the distance calculation, we need to identify the targets with certain matching patterns . What would those patterns be? Does them need to be time point-wised, or individual level?

For this problem we propose a new method to calculate the prediction interval for the

response variable. Do not based on the gamlss prediction but based on a new method to calculate the prediction interval for the response variable. such as the quantile regression methods or other interval estimation methods.

A calibration for the internal validation? We split the training set again into a training set and a validation set. We use the training set to fit the model and the validation set to calibrate interval we get. We can use the residuals from validation set to calibrate the prediction interval for the response variable. Should this be more accurate than the prediction interval based on the training set solely?

Use the weights from the distance, however we can only weight on the individual level. the weights from the time points to calculate the prediction interval for the response variable.

Q2. The prediction interval is not a point-wise prediction interval, but a trajectory-wise prediction interval. What does it mean for the point-wise prediction interval and the trajectory-wise prediction interval?

How to quantify and avoid systematic bias in the predictions?

The point-wise, the trajectory-wise, and the population level prediction interval. we probably need to do a simulation test to see how things going. and Say for example our methods have the same over-estimated coverage rate for the prediction interval, just the same as our simulated results.

- the point-wise coverage probability, which is described as the probability that the interval contains the true value of the response at a given time point. let's say for the 10th day, what is the coverage rate for the prediction interval for the response variable. This is similar to the coverage for the parameters, which is the most commonly used one. The problem rises because we do not have balanced data. we cannot find a same time point for everybody.
- the segmented coverage probability of intervals

piece-wised time segments for coverage rate of the prediction interval. for example, for 10 - 20 days, what is the coverage rate for the prediction interval for the response variable.

But still given each individual or for the population level. what if for certain individual there is no observation in this interval at all.

- the trajectory-wise prediction interval

We can calculate the prediction interval for each trajectory first. The data will be grouped in each individual trajectory, so a proportion of the coverage is calculated for that individual. Then we can calculate the average coverage rate for the whole population.

- the population level prediction interval

This is the methods we currently used, we calculated the predictive interval for each observation whether it is contained in its own predictive interval. Then we take the average of the predictive coverage rate for each observation in each .

as required (Hoerl and Kennard, 1970; Zou and Hastie, 2005). Don't forget to give each section and subsection a unique label (see Sect. **??**).

Paragraph headings. Use paragraph headings as needed.

2.2 *3. Equations*

Here is an equation:

$$f_X(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}; \alpha, \beta, x > 0$$

Here is another:

$$a^2 + b^2 = c^2 \tag{1}$$

Inline equations: $\sum_{i=2}^{\infty}\{\alpha_i^{\beta}\}$

3. Figures and tables

3.1 *Figures coming from R*

Normal figure embedded in text.

```
## Warning in plot.formula(runif(25) ~ runif(25)): the formula 'runif(25) ~
## runif(25)' is treated as 'runif(25) ~ 1'
```

[Figure 1 about here.]

3.2 *Tables coming from R*

```
print(xtable::xtable(head(mtcars)[,1:4],
caption = "Caption centered under table", label = "tab1"),
comment = FALSE, timestamp = FALSE, caption.placement = "top")
```

[Table 1 about here.]

Table 1 shows these numbers. Some of those numbers are plotted in Figure **??**.

References

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* **56,** 337–344.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320.

Received Jan 2024

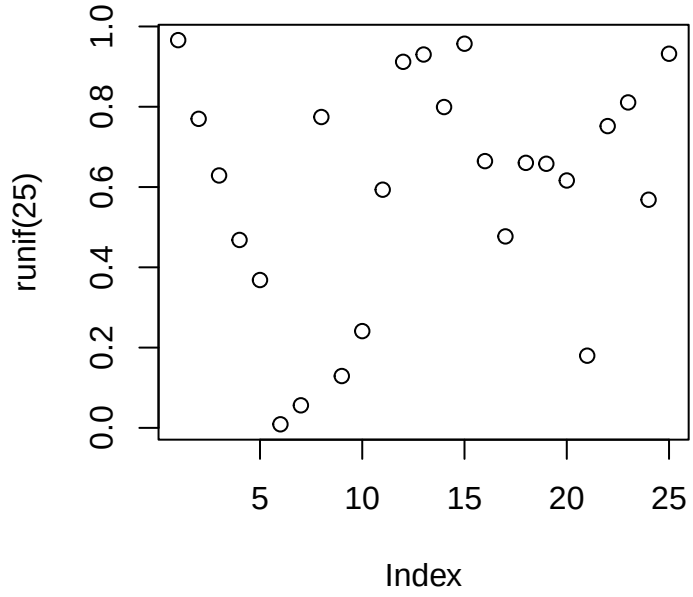

Figure 1. Output from `pdf()`

Table 1
Caption centered under table

	mpg	cyl	disp	hp
Mazda RX4	21.00	6.00	160.00	110.00
Mazda RX4 Wag	21.00	6.00	160.00	110.00
Datsun 710	22.80	4.00	108.00	93.00
Hornet 4 Drive	21.40	6.00	258.00	110.00
Hornet Sportabout	18.70	8.00	360.00	175.00
Valiant	18.10	6.00	225.00	105.00