Study Note on Bühlmann Credibility

1/16/19

The purposes of this Study Note are:

- To provide the traditional Bühlmann-Straub credibility formulas commonly used in insurance,
- To show by example the similarities and differences between the Bühlmann-Straub credibility formulas and the linear random effects model, and
- To briefly introduce the concept of experience rating.

Credibility procedures are in common use in insurance as a way to blend estimates based on group-specific, but thin data, with broader estimates based on more general, but less noisy data. This is essentially the same as a random effects model. While it is beyond the scope of this note, the next logical step, with multi-level groupings, would be to use a hierarchical model.

We will explain this with a numerical example where we estimate random effects within geographic territories (the example would work similarly for other risk groupings), using the following data:

Territory	Year	Risk Count	Average Cost
A	2016	700	$ 900
A	2017	750	$ 800
A	2018	875	$ 1,000
B	2016	350	$ 200
B	2017	400	$ 550
B	2018	425	$ 625
C	2016	100	$ 1,300
C	2017	125	$ 1,800
C	2018	175	$ 2,000
D	2016	675	$ 750
D	2017	700	$ 800
D	2018	725	$ 925
Total		6,000	$ 850

For individual claims y within each group α, the usual actuarial (Bühlmann-Straub) estimators[1] for the expected value of the conditional variance, $\hat{\sigma}^2$ (E(Var(y|α))), and the variance of the conditional expectations, $\hat{\tau}^2$ (Var(E(y|α))), are given by:

$$\hat{\sigma}^2 = E\big(Var(y|\alpha)\big) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n-1}\sum_{j=1}^{n} w_{ij}(X_{ij} - X_i)^2$$

And[2]

$$\hat{\tau}^2 = Var\big(E(y|\alpha)\big) = max\left(0, \frac{1}{U}\left\{\left[\frac{1}{M-1}\sum_{i=1}^{M}\frac{w_i}{W}(X_i - \bar{X})^2\right] - \frac{\hat{\sigma}^2}{W}\right\}\right)$$

where X_{ij} represents the average cost for the i[th] territory in the j[th] year, X_i the risk count weighted average cost for group i across all years, w_{ij} the number of risks within the i[th] territory in the j[th] year, M the number of groups, and n the number of years, and where U is given by:

$$U = \frac{1}{(M-1)}\sum_{i=1}^{M}\frac{w_i}{W}(1 - \frac{w_i}{W})$$

where w_i is the number of risks within the i[th] territory across all years, W is the sum of w_i, and $\bar{X}$ is the average of X_i weighted by w_i. Note that $\hat{\tau}^2$ is an estimate of the variance for an individual risk, so that the model expects that the variance in per-risk average cost for a territory for a year is inversely proportional to the number of risks.

Applying these formulas to the given data yields $\hat{\sigma}^2$= 12,171,436 and $\hat{\tau}^2$= 114,892, and the best linear unbiased prediction of average cost for a given territory in the following year is given by a weighted average of the credibility factor Z_i multiplied by the territory historical average cost X_i, and (1 – Z_i) multiplied by the "complement of credibility," $\hat{\mu}_{new}$, where Z_i is given by $\frac{w_i}{w_i+K}$, K = $\hat{\sigma}^2/\hat{\tau}^2$ = 105.9 (often called the "ballast"), and $\hat{\mu}_{new}$ is given by $\frac{\sum Z_i X_i}{\sum Z_i}$ = $962.45.

[1] The formulas on this page and the following page are adapted from Bühlmann & Gisler, *A Course in Credibility Theory and its Applications*, Springer, 2005, pp. 94-96.

[2] If the maximum is realized, it means there is likely no difference between the groups.

The table of credibilities (Z_i) and best estimates are then given by:

Territory	Risk Count	Average Cost	Credibility	Complement	Best Prediction
A	2,325	$ 905.38	95.6%	$ 962.45	$ 907.86
B	1,175	$ 472.87	91.7%	$ 962.45	$ 513.36
C	400	$ 1,762.50	79.1%	$ 962.45	$ 1,594.98
D	2,100	$ 827.08	95.2%	$ 962.45	$ 833.58
Total	6,000	$ 850.42			$ 850.42

Note that both the historical average cost and the best prediction columns have weighted averages of $850.42. One would have reason to be worried about a process that did not do this. It is important to note, however, that the complement of credibility (which can be interpreted as the expected average cost of a previously unknown employer) is $962.45, NOT $850.42. Also, note that the credibility Z_i is computed based on the number of risks, which, since there are no fixed effects in this model, is proportional to the number of expected claims. It is better both in theory and in practice to use the number of expected claims because in theory, this agrees with linear random effects models, but in practice, one needs to recognize the superior experience of a policyholder who never has a claim, whereas giving them zero credibility treats them like an average policyholder.

Now, let us compare this with using the lmer function in the R package lme4[3]:

```
model1 <- lmer(averagecost ~ 1 + (1 | territory), data = territorydata,
          weights = riskcount)
```

This is how one specifies a random intercepts model in lme4.

Let's look at the basic output:

```
summary(model1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: averagecost ~ 1 + (1 | territory)
   Data: territorydata
Weights: riskcount

REML criterion at convergence: 157.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5553 -0.6536  0.1390  0.7386  1.2030
```

[3] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker, Fitting Linear Mixed-Effects Models Using lme4, *Journal of Statistical Software*, 67(1), 2015, 1-48. doi:10.18637/jss.v067.i01.

```
Random effects:
 Groups    Name        Variance Std.Dev.
 territory (Intercept)   264378  514.2
 Residual              12351241 3514.4
Number of obs: 12, groups:  territory, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)    977.4      263.2   3.713
```

The estimate of the residual variance (12,351,241) is similar to that from the Bühlmann-Straub formulas (12,171,436). The REML procedure and the Bühlmann-Straub procedure use different estimators, though they are estimating the same model. The estimate of the fixed effect for the intercept (977.40) corresponds to the estimate of $\hat{\mu}_{new}$ in the Bühlmann-Straub setup (962.45). While these estimates are close, they do not agree exactly because the estimators used are different.

Note that the model object does not explicitly provide the random effect estimates that give best linear unbiased predictions, but these are available via the ranef function:

```
ranef(model1)
```

```
$`territory`
  (Intercept)
A   -70.63284
B  -485.26205
C   702.96812
D  -147.07323
```

Adding these to the intercept of 977.4 yields lmer's Best Linear Unbiased Predictions (BLUPs) for each territory:

Territory	Best Linear Unbiased Prediction
A	906.77
B	492.14
C	1,680.37
D	830.33

Note that the output of ranef could easily be reproduced with just the information in the model summary. This is because the BLUPs depend only on the variance components and on the fixed effect intercept, and they depend on them in precisely the same way the BLUPs depend on $\hat{\sigma}^2$, $\hat{\tau}^2$, and $\hat{\mu}_{new}$. The estimates of the variance components are the only difference between the processes. (The fixed-effects intercept is different from $\hat{\mu}_{new}$ only because each is calibrated to make the weighted average best predictions equal the weighted average historical costs.) In fact, we can reproduce the previous table using lmer instead of Bühlmann-Straub:

Territory	Risk Count	Average Cost	Credibility	Complement	Best Prediction
A	2,325	$ 905.38	98.0%	$ 977.43	$ 906.80
B	1,175	$ 472.87	96.2%	$ 977.43	$ 492.17
C	400	$ 1,762.50	89.5%	$ 977.43	$ 1,680.40
D	2,100	$ 827.08	97.8%	$ 977.43	$ 830.36
Total	6,000	$ 850.42			$ 850.42

Also note that only the ratio K of variance components and the complement of credibility matter, at the end of the day. Everything else is simple arithmetic. Insurance professionals sometimes choose a value of K based on long experience across multiple datasets and use it across multiple similar problems without re-estimating it. The underlying motivation is that variance estimates are noisy, and it's useful to include as much information as possible in arriving at K, even beyond the dataset under consideration. On occasion, insurance professionals also occasionally modify the complement of credibility to take into account knowledge that is not contained in the data. These can be reasonable thing to do, but must be appropriately documented as they may have impacts on the estimates that the audience for the results of the analysis might not expect. In particular, any change to the complement will cause the historical average cost and the average best prediction to differ and should be carefully noted.

Experience Rating

Experience rating refers to the practice of adjusting the manual renewal premium, calculated based on policy characteristics according to the rating plan, by an individual experience modification factor representing the degree to which actual past claims experience may be a credible predictor of future claims experience. For example, in workers' compensation insurance, the manual typically states a rate per $100 of payroll for each class of employee (the rate being very different for road construction workers vs. clerical workers). Experience rating then treats each employer in the way each territory was treated in the above example, to arrive at an adjusted rate for that employer. (The actual details are more complex, as credibility weighting is applied separately for different loss layers, giving less credibility for higher layers, since fewer claims are expected to pierce those layers—at the end of the day the weights w_i should be proportional to the expected number of non-zero claims, which will be fewer for higher layers .). Thus, experience rating allows insurance companies to recognize indirect, or unobservable, risk characteristics, such as unobservable features of workplace safety practices at medium and large-size employers. Note that typically there is more variability between, and less variability within, individual policyholders. This means K is usually considerably smaller in an experience rating model as compared to a credibility model for territories, sometimes by an order of magnitude. For example, an employer with $100,000 of workers compensation manual premium (i.e., premium prior to the application of the experience mod factor) may have experience-rating credibility comparable to the territorial ratemaking credibility of a territory with $1,000,000 of manual premium.