Bootstrapping the triangles

Prečo, kedy a ako (NE)bootstrapovať v trojuholníkoch?

Michal Pešta

Charles University in Prague

Faculty of Mathematics and Physics

Actuarial Seminar, Prague

19 October 2012

Overview

- Idea and goal of bootstrapping
- Bootstrap in reserving triangles
- Residuals
- Diagnostics
- Real data example

- Support:

  Czech Science Foundation project "DYME Dynamic Models in Economics" No. P402/12/G097

Prologue for Bootstrap

- computationally intensive method popularized in 1980s due to the introduction of computers in statistical practice
- a strong mathematical background ⇝ bootstrap does not replace or add to the original data
- unfortunately, the name "bootstrap" conveys the impression of "something for nothing" ⇝ idly resampling from their samples

Reserving Issue

- consider traditional actuarial approach to reserving risk ... the uncertainty in the outcomes over the lifetime of the liabilities
- bootstrap can be also applied under Solvency II ... outstanding liabilities after 1 year
- distribution-free methods (e.g., chain ladder) only provide a standard deviation of the ultimates/reserves (or claims development result/run-off result)

? another risk measure (e.g., VaR @ 99.5%)

↯ moreover, distributions of ultimate cost of claims and the associated cash flows (not just a standard deviation)?

! claims reserving technique applied mechanically and without judgement

Bootstrap

- simple (distribution-independent) resampling method
- estimate properties (distribution) of an estimator by sampling from an approximating (e.g., empirical) distribution
- useful when the theoretical distribution of a statistic of interest is complicated or unknown

Resampling with replacement

- random sampling with replacement from the original dataset $\rightsquigarrow$ for $b = 1, \ldots, B$ resample from $X_1, \ldots, X_n$ with replacement and obtain $X^*_{1,b}, \ldots, X^*_{n,b}$

Bootstrap example

- input data (# of catastrophic claims per year in 10y history): $35, 34, 13, 33, 27, 30, 19, 31, 10, 33 \rightsquigarrow$ $mean = 26.5,\ sd = 9.168182$

Case sampling

- bootstrap sample 1 (1st draw with replacement): $30, 27, 35, 35, 13, 35, 33, 34, 35, 33 \rightsquigarrow$ $mean_1^* = 31.0,\ sd_1^* = 6.847546$

$\vdots$

- bootstrap sample 1000 (1000th draw with replacement): $19, 19, 31, 19, 33, 34, 31, 34, 34, 10 \rightsquigarrow$ $mean_{1000}^* = 26.4,\ sd_{1000}^* = 8.771165$

- $mean_1^*, \ldots, mean_{1000}^*$ provide bootstrap empirical distribution for $mean$ and $sd_1^*, \ldots, sd_{1000}^*$ provide bootstrap empirical distribution for sd (REALLY!?)

Terminology

- $X_{i,j}$ ... claim amounts in development year j with accident year i
- $X_{i,j}$ stands for the incremental claims in accident year i made in accounting year $i+j$
- n ... current year – corresponds to the most recent accident year and development period
- Our data history consists of right-angled isosceles triangles $X_{i,j}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, n+1-i$

Run-off (incremental) triangle

Accident year i	Development year j				
	1	2	$\cdots$	$n-1$	n
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,n-1}$	$X_{1,n}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,n-1}$	
			$\ddots$		
$\vdots$	$\vdots$	$\vdots$	$X_{i,n+1-i}$		
$n-1$	$X_{n-1,1}$	$X_{n-1,2}$			
n	$X_{n,1}$				

Notation

- $C_{i,j}$ ... cumulative payments in origin year i after j development periods

$$C_{i,j} = \sum_{k=1}^{j} X_{i,k}$$

- $C_{i,j}$ ... a random variable of which we have an observation if $i + j \leq n + 1$
- Aim is to estimate the ultimate claims amount $C_{i,n}$ and the outstanding claims reserve

$$R_i = C_{i,n} - C_{i,n+1-i}, \quad i = 2, \ldots, n$$

by completing the triangle into a square

Run-off (cumulative) triangle

Accident year i	Development year j				
	1	2	$\cdots$	$n-1$	n
1	$C_{1,1}$	$C_{1,2}$	$\cdots$	$C_{1,n-1}$	$C_{1,n}$
2	$C_{2,1}$	$C_{2,2}$	$\cdots$	$C_{2,n-1}$	
			$\ddots$		
$\vdots$	$\vdots$	$\vdots$	$C_{i,n+1-i}$		
$n-1$	$C_{n-1,1}$	$C_{n-1,2}$			
n	$C_{n,1}$				

Theory Behind the bootstrap

- validity of bootstrap procedure
- asymptotically distributionally coincide

$$\widehat{R}_i^* - \widehat{R}_i \Big| \{X_{i,j} : i + j \leq n + 1\} \quad \text{and} \quad \widehat{R}_i - R_i$$

- approaching (each other) in distribution in probability along $D_n^{(n)} = \{X_{i,j} : i + j \leq n + 1\}$

$$\boxed{\widehat{R}_i^* - \widehat{R}_i \Big| D_n^{(n)} \overset{\mathsf{D}}{\longleftrightarrow} \widehat{R}_i - R_i \quad \text{in probability}, \quad n \to \infty}$$

- $\forall$ real-valued bounded continuous function f

$$\mathsf{E}\left[f\left(\widehat{R}_i^* - \widehat{R}_i\right)\Big| D_n^{(n)}\right] - \mathsf{E}\left[f\left(\widehat{R}_i - R_i\right)\right] \xrightarrow[n\to\infty]{\mathsf{P}} 0$$

Residuals

- measure the discrepancy of fit in a model
- can be used to explore the adequacy of fit of a model
- may also indicate the presence of anomalous values requiring further investigation
- in regression type problems, it is common to bootstrap the residuals, rather than bootstrap the data themselves
- should mimic iid r.v. having zero mean, common variance, symmetric distribution

Raw residuals

$$_{(R)}r_{i,j} = X_{i,j} - \widehat{X}_{i,j} \quad \text{or} \quad _{(R)}r_{i,j} = C_{i,j} - \widehat{C}_{i,j}$$

- in homoscedastic linear regression, CL with $\alpha = 0$, GLM with normal distribution and identity link

Pearson residuals

$$_{(P)}r_{i,j} = \frac{X_{i,j} - \widehat{X}_{i,j}}{\sqrt{V(\widehat{X}_{i,j})}}$$

- in GLM or GEE
- idea is to <u>standardize</u> raw residuals
- ODP:

$$_{(P)}r_{i,j} = \frac{X_{i,j} - \widehat{X}_{i,j}}{\sqrt{\widehat{X}_{i,j}}}$$

- Gamma GLM:

$$_{(P)}r_{i,j} = \frac{X_{i,j} - \widehat{X}_{i,j}}{\widehat{X}_{i,j}}$$

Anscombe residuals I

$$ {}_{(A)}r_{i,j} = \frac{A(X_{i,j}) - A(\widehat{X}_{i,j})}{A'(\widehat{X}_{i,j})\sqrt{V(\widehat{X}_{i,j})}} $$

- in GLM or GEE
- disadvantage of Pearson residuals: often markedly skewed
- idea is to obtain residuals, which are <u>not skewed</u> (to "normalize" residuals)

Anscombe residuals II

$$A(\cdot) = \int \frac{\mathsf{d}\mu}{V^{1/3}(\mu)}$$

- ODP:

$$_{(A)}r_{i,j} = \frac{3}{2}\frac{X_{i,j}^{2/3} - \widehat{X}_{i,j}^{2/3}}{\widehat{X}_{i,j}^{1/6}}$$

- Gamma GLM:

$$_{(A)}r_{i,j} = 3\frac{X_{i,j}^{1/3} - \widehat{X}_{i,j}^{1/3}}{\widehat{X}_{i,j}^{1/3}}$$

- inverse Gaussian:

$$_{(A)}r_{i,j} = \frac{\log X_{i,j} - \log \widehat{X}_{i,j}}{\widehat{X}_{i,j}^{1/2}}$$

Deviance residuals

$$_{(D)}r_{i,j} = sign(X_{i,j} - \widehat{X}_{i,j})\sqrt{d_i},$$

where d_i is the deviance for one unit, i.e., $\sum d_i = D$

- in GLM or GEE
- similar advantageous properties like Anscombe residuals (see Taylor expansion)
- ODP:

$$_{(D)}r_{i,j} = sign(X_{i,j} - \widehat{X}_{i,j})\sqrt{2\left(X_{i,j}\log(X_{i,j}/\widehat{X}_{i,j}) - X_{i,j} + \widehat{X}_{i,j}\right)}$$

Scaled residuals

- scaling parameter ϕ

$$^{(sc)}r_{i,j} = \frac{r_{i,j}}{\sqrt{\widehat{\phi}}}$$

- to obtain the bootstrap prediction error, it is necessary to add an estimate of the process variance ... bias correction in the bootstrap estimation variance
- Pearson scaled residuals in ODP for the bootstrap prediction error:

$$^{(sc)}_{(P)}r'_{i,j} = {}_{(P)}r_{i,j} \times \left(\frac{\sum_{i,j \leq n+1} {}_{(P)}r^2_{i,j}}{n(n+1)/2 - (2n+1)} \right)^{-1/2}$$

Adjusted residuals

- alternative to scaled residuals ... bias correction in the bootstrap estimation variance

$$^{(adj)}r_{i,j} = r_{i,j} \times \sqrt{\frac{n-p}{n}}$$

Diagnostics for residuals

- residuals and squared residuals
- type of plots:
  - accident years
  - accounting years
  - development years
- no pattern visible
- iid with zero mean, symmetrically distributed, common variance

Bootstrap procedure I

- Ex: bootstrap in ODP

(1) fit a model $\rightsquigarrow$ estimates $\widehat{\alpha}_i, \widehat{\beta}_j, \widehat{\gamma}, \widehat{\phi}$

(2) fitted (expected) values for triangle
$\widehat{X}_{i,j} = \exp\{\widehat{\gamma} + \widehat{\alpha}_i + \widehat{\beta}_j\}$

(3) scaled Pearson residuals

$$^{(sc)}_{(P)}r_{i,j} = \frac{X_{i,j} - \widehat{X}_{i,j}}{\sqrt{\widehat{\phi}\widehat{X}_{i,j}}}$$

Bootstrap procedure II

(4) resample residuals $\left\{ {}^{(sc)}_{(P)} r_{i,j} \right\}$ B-times with replacement $\rightsquigarrow$ B triangles of bootstrapped residuals $\left\{ {}^{(sc)}_{(P,b)} r^{*}_{i,j} \right\}$, $1 \leq b \leq B$

(5) construct B bootstrap triangles

$$
{}_{(b)}X^{*}_{i,j} = {}^{(sc)}_{(P,b)} r^{*}_{i,j} \sqrt{\widehat{\phi} \widehat{X}_{i,j}} + \widehat{X}_{i,j}
$$

(6) perform ODP on each bootstrap triangle $\rightsquigarrow$ bootstrapped estimates ${}_{(b)}\widehat{\alpha}_i, {}_{(b)}\widehat{\beta}_j, {}_{(b)}\widehat{\gamma}, {}_{(b)}\widehat{\phi}$

Bootstrap procedure III

(7) calculate bootstrap reserves

$$_{(b)}\widehat{R}_i^* = \sum_{j=n+2-i}^{n} {}_{(b)}\widehat{X}_{i,j}^* = \exp\{_{(b)}\widehat{\gamma} + {}_{(b)}\widehat{\alpha}_i\} \sum_{j=n+2-i}^{n} \exp\{_{(b)}\widehat{\beta}_j\}$$

- (4)–(7) is a bootstrap loop (repeated B-times)

(8) empirical distribution of size B for the reserves $\rightsquigarrow$ empirical (estimated) mean, s.e., quantiles, ...

Taylor and Ashe (1983) data

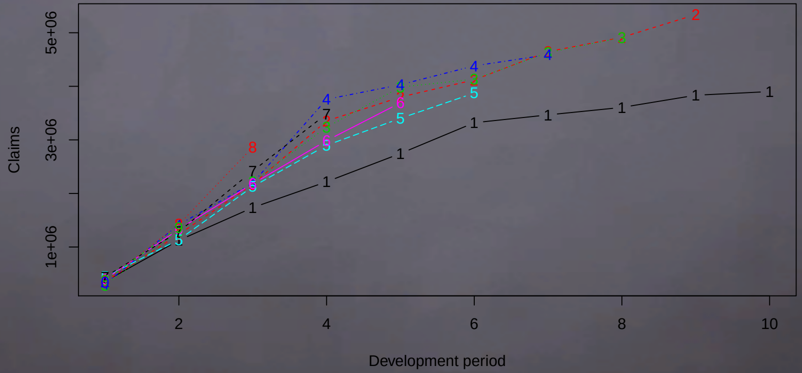
- incremental triangle

357 848	766 940	610 542	482 940	527 326	574 398	146 342	139 950
352 118	884 021	933 894	1 183 289	445 745	320 996	527 804	266 172
290 507	1 001 799	926 219	1 016 654	750 816	146 923	495 992	280 405
310 608	1 108 250	776 189	1 562 400	272 482	352 053	206 286	
443 160	693 190	991 983	769 488	504 851	470 639		
396 132	937 085	847 498	805 037	705 960			
440 832	847 631	1 131 398	1 063 269				
359 480	1 061 648	1 443 370					
376 686	986 608						
344 014							

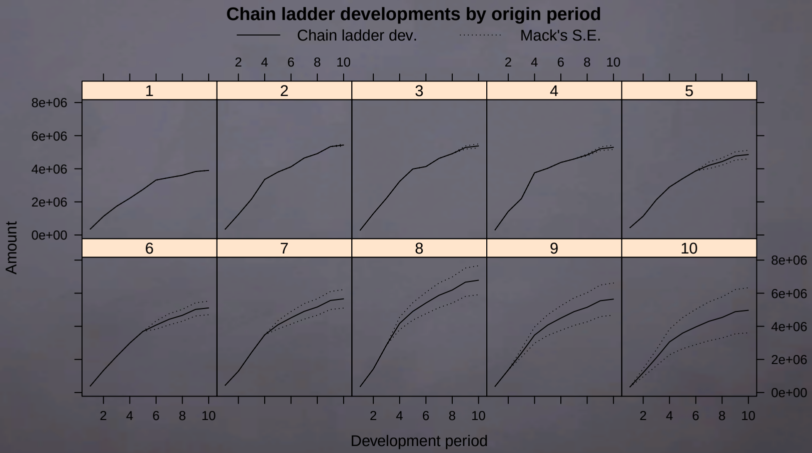
227 229	67 948
425 046	

- R software, ChainLadder package

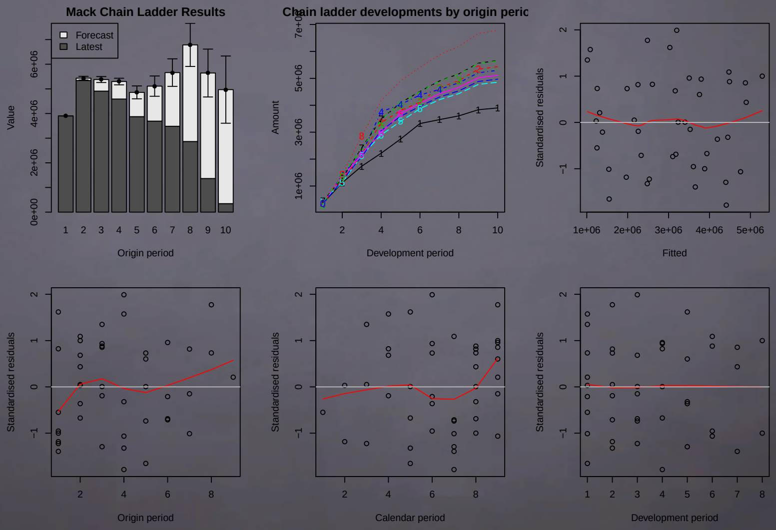
Development of claims



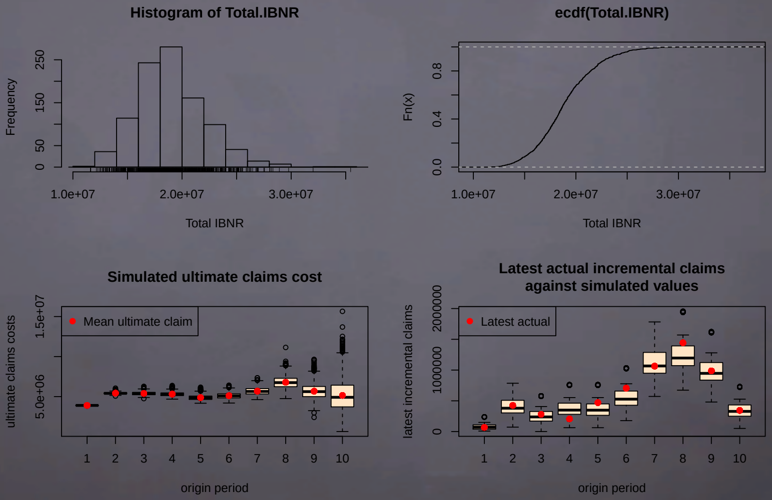
CL with Mack's s.e.



Chain ladder diagnostics



Bootstrap results



Mack CL vs bootstrap GLM (ODP)

Accident	Chain Ladder			Bootstrap		
year	Ultimate	IBNR	S.E.	Ultimate	IBNR	S.E.
1	3 901 463	0	0	3 901 463	0	0
2	5 433 719	94 634	75 535	5 434 680	95 595	106 313
3	5 378 826	469 511	121 699	5 396 815	487 500	222 001
4	5 297 906	709 638	133 549	5 315 089	726 821	265 696
5	4 858 200	984 889	261 406	4 875 837	1 002 526	313 015
6	5 111 171	1 419 459	411 010	5 113 745	1 422 033	377 703
7	5 660 771	2 177 641	558 317	5 686 423	2 203 293	487 891
8	6 784 799	3 920 301	875 328	6 790 462	3 925 964	789 329
9	5 642 266	4 278 972	971 258	5 675 167	4 311 873	1 034 465
10	4 969 825	4 625 811	1 363 155	5 148 456	4 804 442	2 091 629
Total	53 038 946	18 680 856	2 447 095	53 338 139	18 980 049	3 096 767

Comparison of distributional properties

- why to bootstrap ?
- moment characteristics (mean, s.e., ...) does not provide full information about the reserves' distribution
- additional assumption required in the classical approach
- 99.5% quantile necessary for VaR
  - assuming normally distributed reserves ... 24 984 154
  - assuming log-normally distributed reserves ... 25 919 050
  - bootstrap ... 28 201 572

Conclusions

- mean and variance do not contain full information about the distribution $\rightsquigarrow$ cannot provide quantities like VaR
- assumption of log-normally distributed claims $\not\Leftrightarrow$ log-normally distributed reserves (far more restrictive)
- various types of residuals
- bootstrap (simulated) distribution mimics the unknown distribution of reserves (a mathematical proof necessary)
- R software provides a free sufficient actuarial environment for reserving

References

- England, P. D. and Verrall, R. J. (1999)
  Analytic and bootstrap estimates of prediction errors in claims reserving.
  Insurance: Mathematics and Economics, 25.
- England, P. D. and Verrall, R. J. (2006)
  Predictive distributions of outstanding claims in general insurance.
  Annals of Actuarial Science, 1 (2).
- Pesta, M. (2012)
  Total least squares and bootstrapping with application in calibration.
  Statistics: A Journal of Theoretical and Applied Statistics.

Thank you !

pesta@karlin.mff.cuni.cz