Rapp

Non-life tariff analysis in practice with Rapp

Lecture in Indonesia August 2015

Stig Rosenlund

The methods of this talk are implemented in the programming language Rapp. To find Rapp on the internet, search **free actuarial language.** It is presently in a review process to become an Open Source program.

About me: PhD in Mathematical Statistics at the University of Göteborg in 1983. Actuary at the Swedish insurance company Länsförsäkringar 1983-2010. After retirement in 2010, I have resumed my research, now focusing on insurance mathematics. So far I have in the 2010's had three papers published - two in tariff analysis and one in reserving.

Please interrupt me with questions at any time!

2015, Stig Rosenlund

What is tariff analysis?

It is to find the best estimate of the pure premium = expected claim cost of insuring an object during one year. The most common form of a tariff is a collection of factors for different arguments = background variables. For instance, the insurance premium of your car might be obtained by multiplying a base factor, eg. 200,000 Rupiah, with a factor 1.10 due to your age, a factor 0.80 due to the car age and a factor 1.20 due to the region where the car is used. In all 211,000 Rupiah.

I will base this lecture on my paper

Rosenlund, S. (2014).
Inference in multiplicative pricing.
Scandinavian Actuarial Journal **2014**(8), 690-713.

Tariff analysis point estimate methods

MMT – Method of Marginal Totals

Model: Risk premium is multiplicative in the arguments.

Method: Solve a system of equations defined by prescribing that the sum of multiplicatively computed estimated claim costs over any argument class - any marginal - be equal to the empirical claim cost of the argument class.

Standard GLM

Model: Claim numbers are Poisson distributed. Claim severities are Γ distributed with a constant CV (coefficient of variation).

Method: Solve the ML (maximum likelihood) equations resulting from the model.

The literature also treats the Tweedie method for risk premium, but I have shown that it should not be used.

Tariff analysis variance estimate methods

MVW – MMT Variance estimates under Weak assumptions

Model: Claim frequency and mean claim are multiplicative in the arguments. The claim cost of a tariff cell is distributed Compound Poisson.

Method: The GLM Poisson log link model for claim numbers is used for claim frequency. For mean claim I take the estimated CVs for univariate mean claims as a start values. These are adjusted upwards by factors resembling the ratios of GLM claim frequency variance estimates to simple univariate claim frequency variance estimates.

Standard GLM

Model: As before.
Method: The variance estimates that follow from the GLM theory.

The following are my main conclusions.

1. With sufficiently many claims *or* sufficiently many arguments MMT is preferable over Standard GLM in tariff analysis.

2. The MVW confidence interval method for MMT is mostly preferable to Standard GLM.

To show or disprove conclusion **1**, I had to define preferable by means of a measure of goodness-of-fit for a method. I proposed the exposure-weighted mean square deviation of estimated risk premium from true risk premium, summed over all tariff cells. Thus, let u be the index of a tariff cell $\in \{1, 2, \ldots, n\}$ and

e_u = exposure in cell u
τ_u = true risk premium for cell u
$\widehat{\tau}_u^{(\mathrm{X})}$ = estimate of τ_u for a method X

and define the goodness-of-fit measure

$$\mathcal{M}(\mathrm{X}, \{e_u\}) = \mathrm{E}\left[\sum_{u=1}^{n} e_u \left(\widehat{\tau}_u^{(\mathrm{X})} - \tau_u\right)^2\right] / \sum_{u=1}^{n} e_u$$

Now introduce a volume measure c such that

$$e_u = ce_u^0$$

for some suitably normed sequence $\{e_u^0\}$. The interest is in how the goodness-of-fit measure for X, compared to some other method, behaves for different values of c.

Thus I defined

$$\begin{aligned} \mathcal{M}_M(c) &= \mathcal{M}(\text{MMT}, \{ce_u^0\}) \\ \mathcal{M}_S(c) &= \mathcal{M}(\text{Standard GLM}, \{ce_u^0\}) \end{aligned}$$

Consider a random variable Z with mean μ that we want to be close to a number a. It holds

$$\text{E}[(Z-a)^2] = \text{Var}[Z] + (a-\mu)^2$$

or in words

$$\text{Mean square error} = \text{Variance} + \text{bias}^2$$

We can generalize this to the collection of all tariff cells.

I and other authors have found indications that normally, vaguely speaking, Standard GLM has smaller variance than MMT. Both methods have biases. I

conjectured in my article that variances, suitably defined for the collection of all tariff cells, k_1/c for MMT and k_2/c for Standard GLM, where $k_1 > k_2$, hold asymptotically for both $c \to 0$ and $c \to \infty$, albeit possibly with one pair (k_1, k_2) for $c \to 0$ and another pair (k_1, k_2) for $c \to \infty$. This conjecture was corroborated by my simulations. The means that when $c <$ some constant, then Standard GLM will be preferable.

For the biases, I had to distinguish between

A. The true risk premiums are exactly multiplicative.

B. The true risk premiums deviate from exact multiplicativity.

In case **A** both methods will have asymptotic zero bias as $c \to \infty$. If my conjecture is correct, then Standard GLM will be preferable both as $c \to 0$ and as $c \to \infty$.

On the other hand, if **B** holds, which almost always will be the case for more than one argument, then the comparison of asymptotic goodness-of-fit for the two methods will be determined by the now non-zero asymptotic biases, suitably defined.

I conjectured that in case B the asymptotic bias of MMT is typically smaller than the asymptotic bias of Standard GLM. This was also corroborated by my simulated cases. These biases could be determined exactly by taking cases were all observed tariff risk premiums were exactly equal to their expected values.

Thus, since the variances tend to 0, I found for my cases

$$\lim_{c\to\infty} \mathcal{M}_M(c) < \lim_{c\to\infty} \mathcal{M}_S(c)$$

There are exceptions, but almost always there should be an indifference value c_0 of c such that

- $\mathcal{M}_M(c) > \mathcal{M}_S(c)$ for $c < c_0$ — Standard GLM is preferable.
- $\mathcal{M}_M(c_0) = \mathcal{M}_S(c_0)$ — Standard GLM and MMT are equal.
- $\mathcal{M}_M(c) < \mathcal{M}_S(c)$ for $c > c_0$ — MMT is preferable.

I could also show that the more arguments there are, the smaller c_0 is. An illustration is given on the next page.

I leave the demonstration of conclusion **2** for confidence intervals to another occasion.

Rapp

Mean square error MSE depending on exposure for MMT and Standard GLM

Practical tariff analysis

My focus will now be on the practical steps of insurance tariff analysis using a multiplicative GLM model, by which I also can mean MMT.

① Which data are required from the production system?

② How do you prepare data to be input to GLM equation solving?

③ How to present analyses to the managers of business lines?

Practical tariff analysis involves much contact with database managers of the insurance company. You have to communicate your needs clearly. IT personnel has to be assigned to the task of regularly suppling data to the data warehouse used by the actuaries.

Much effort must be spent on securing reliable and relevant data bases with claim payments and changes of claim-handler reserves, so that claim reserves (= estimates of remaining payment sums) can be computed.
I will treat this aspect lightly, just assuming that claim reserves are registered in the individual claims in the claim data warehouse.

1. Which data?

With the best-estimate purpose in mind, we want to have data giving the claim cost of any customer for some years back and giving all possible relevant background info on any customer, subject to the limitations of law. We want to be able to tie all such data to the time-points when they were valid.

Claims

Firstly what you do NOT need in the claims data warehouse. You do not need background info that is registered in the insurance policies data base. Do not burden the IT department with requests to have such info in the claims. You will find later that some important info has been forgotten.

Instead: secure the availability of reliable such info in the insurance data warehouse, and secure the capacity to transfer the info to the claims. Transfer how? I will tell you later.

Claim data warehouse structure

Often in production systems there are hierarchical layers of claim tables with main claims and parts of these. For instance a main claim for house damage with parts for damage from fire, water, etc. I will simplify here

and just consider one part in the claim.

Store one line per claim in the actuarial claim table of the data warehouse. You need a variable for paid and a variable for reserve. If the reserve variable is the claim-handler reserve, and if this typically undershoots or overshoots the actual remaining payments, then the actuary will have to compute both IBNR (Incurred But Not Reported) and RBNS (Reported But Not Settled) reserves to get a percentage add-on for total claim cost. At my former employer Länsförsäkringar the claim-handler reserves typically overshoot, due to a safety principle.

Do not spend much time on systems for regularly aggregating claims on some background variables. When doing tariff analysis you will invariably find that these background variables are not sufficient.

Needed claim variables

- ✔ Unique identity variables of the insurance policy hit by the claim. For Indonesian motor insurance these are supposed to be Kode_Perusahaan and Nomor_Rangka (Company and ChassisID).
- ✔ Cover used by the claim, eg. theft.
- ✔ Claim occurrence date.
- ✔ Claim reporting date.
- ✔ Claim settlement date - 0 if claim is open.
- ✔ Paid.
- ✔ Reserve.
- ✔ Amount deducted.

Rapp

Insurance policies

You need policy *periods*, ie. one line per insured period with the date it begun and the date it was succeeded by a new period or canceled. This gives you complete freedom to calculate exposure and claim cost for any period of time for which you have all periods in effect in your data warehouse.

Determine a first date, such that you have access to all periods in effect from that date and such that it is sufficiently close in time to make those periods relevant. The larger portfolio, the more recent you can make that date.

Often such periods will overlap, since a suitable insurance production

data base will keep records unchanged to not destroy information. For instance, if a policy was initiated for one year but the customer got a change in cover after half a year, then you have two periods which overlap. Thus you need to shorten the first period before computing exposure.

Policy period data warehouse structure

Often in production systems there are hierarchical layers of insurance policy tables with a main table and tables for different covers under this. I will simplify here and just consider one policy table with covers stored in a finite number of variables side by side.

Store one line per policy period in the actuarial data warehouse.

Do not spend too much time on systems for regularly aggregating policy periods on some background variables.

Needed policy period variables

- ✔ Unique identity variables of the insurance policy, eg. Kode_Perusahaan and Nomor_Rangka.
- ✔ Beginning date, which I will call FromDate.
- ✔ Date after last day the period was valid, which I will call UntoDate.
- ✔ Cover(s) of the policy period, which can be registered in many ways.
- ✔ All possibly relevant background info stored in a way to make possible exact computation. For instance, store the birth date of the owner rather than her age at some point in the period.

Possibly needed policy period variable

✔ Code for validity, so that never valid periods can be excluded.

Useful policy period variables

✔ Premiums and taxes in the form of yearly or monthly rates.

Technical format of the data warehouse

Use a form that lets you display data on the screen easily. SAS tables are possible. Access tables would be cheaper. Excel tables are awkward for large files. Rapp uses plain text files exported from SAS or Access.

2. Steps of tariff analysis with Rapp

Many actuaries use SAS to mangle data into the right form for tariff

analysis and subsequently do a GLM tariff analysis. I have done so myself when I was employed at Länsförsäkringar.

Instead of SAS you can use the programming language Rapp, which I have developed for actuarial analyses. Rapp can do both data mangling and GLM equation solving. It uses a SAS-like syntax. It is free.

Rapp is not a full-fledged language, since it does not have if/else-, goto- and loop-statements like SAS. But it has some facilities especially designed for actuarial needs. I will show you shortly.

Rapp executes fast in data mangling, although not always faster than SAS. It is very fast in GLM equation solving - many times faster than SAS Proc

Genmod. Also it is very flexible.

So these are the steps. Selected files are tab-delimited plain text files.

- Select claims, one per line. Do not bother to aggregate data. You can put claims for different covers, such as Third Party Liability and Property damage, in the same file. Rapp can make subselections in the GLM equation solving program.
- Select policy periods, one per line. Do not bother to aggregate data. Make some initial manglings, like deducing simpler cover variables from complex ones in the data warehouse.
- Eliminate policy period overlaps. Here is a Rapp proc for it. At the same time it computes the insured's age at the beginning of the period.

```
Proc Ovelim run
  Infil("C:\Rapp\Data\temp1.Txt") dlm(9)
  var(
    Kode_Perusahaan $ Nomor_Rangka $ FromDate UntoDate
    Cover1 $ Cover2 $ Backgroundvar1 $ Backgroundvar2 $
    BirthDate Yearly_premium R
  )
  Dvar( Age = min(99,|[(FromDate-BirthDate)/10000]|) )
  FromDatevar(FromDate) UntoDatevar(UntoDate)
  key(Kode_Perusahaan Nomor_Rangka)
  Utfil(C:\Rapp\Data\temp2.Txt)
Endproc
```

This shortens UntoDate to at most the value of the next policy period's FromDate, if a next policy period with the same Kode_Perusahaan, Nomor_Rangka exists.

➥ Transfer insurance info to claims. You take the policy period variables from the period with the same identity variables and the largest FromDate that is at most claim occurrence date. A Rapp proc for it:

```
Proc Match unmatch(t) umvalue(0) stats ;
  Masterfil fil(C:\Rapp\Data\temp2.Txt) dlm(9)
    var(
      Kode_Perusahaan $ Nomor_Rangka $ FromDate UntoDate
      Cover1 $ Cover2 $ Backgroundvar1 $ Backgroundvar2 $
      BirthDate Yearly_premium R Age
    )
    Key(Kode_Perusahaan Nomor_Rangka) Timekey(FromDate)
  ;
  Transfil fil(C:\Rapp\Data\tempc1.Txt) dlm(9)
    var(
      Kode_Perusahaan $ Nomor_Rangka $ ClaimDate ReportDate
      SettlementDate CoverUsed $ Paid R Reserve R Deducted R
    )
    Key(Kode_Perusahaan Nomor_Rangka) Timekey(ClaimDate)
  ;
  Utfil fil(C:\Rapp\Data\Claims.Txt) dlm(9) noq
    var(
      Kode_Perusahaan $ Nomor_Rangka $ FromDate UntoDate
      Cover1 $ Cover2 $ Backgroundvar1 $ Backgroundvar2 $
      Age Yearly_premium R ClaimDate ReportDate SettlementDate
      CoverUsed $ Paid R Reserve R Deducted R
    )
  ;
Endproc
```

➥ Compute exposure, ie. let Rapp — for each year you want an account of — compute the duration between 0 and 1 for each policy period. Example. It adds three new variables: Period name 2009, duration and earned premium in [20090101,20091231]. Likewise for 2010-2013.

```
Proc Durber infil(C:\Rapp\Data\temp2.Txt)
  utfil(C:\Rapp\Data\Idur.Txt)
  var(
    Kode_Perusahaan $ Nomor_Rangka $ FromDate UntoDate
    Cover1 $ Cover2 $ Backgroundvar1 $ Backgroundvar2 $
    BirthDate Yearly_premium R Age
  )
  Frdkvar(FromDate) Todkvar(UntoDate)
  Ypremvar(Yearly_premium)
  Datum(
    20090101 20091231 2009
    20100101 20101231 2010
    20110101 20111231 2011
    20120101 20121231 2012
    20130101 20131231 2013
  )
Endproc
```

➥ Do a tariff analysis.

```
Include C:\Rapp\Rpp\Init.Rpp
Proc Init lan(e) Endproc
Proc Taran listfil(C:\Rapp\Txt\T01.Txt) rub62 'Cover 1';
Infiler fil(C:\Rapp\Data\Idur.Txt) dlm(9)
  Var(
    Kode_Perusahaan $ Nomor_Rangka $ FromDate UntoDate
    Cover1 $ Cover2 $ Backgroundvar1 $ Backgroundvar2 $
    BirthDate Yearly_premium R Age Pername Dur R Prem R
  )
  Urval(Cover1 = 'Y')
;
Infiler fil(C:\Rapp\Data\Claims.Txt) dlm(9)
  Var(
    Kode_Perusahaan $ Nomor_Rangka $ FromDate UntoDate
    Cover1 $ Cover2 $ Backgroundvar1 $ Backgroundvar2 $
    Age Yearly_premium R ClaimDate ReportDate SettlementDate
    CoverUsed $ Paid R Reserve R Deducted R
  )
  Dvar(
    Pername = [ClaimDate/10000]
    Skkost = Paid+Reserve+Deducted
  )
```

```
  Urval(Cover1 = 'Y' & Coverused = '1'
    & ClaimDate >= 20090101 & ClaimDate <= 20131231)
;
arg(Pername) rub110 'Calendar year';
arg(Backgroundvar1) rub110 'Car make';
arg(Backgroundvar2) rub110 'Region' tar(0.7 0.8 1 1.2) ;
arg(Age) rub110 'Age of insured'
  niv( // Format (, lower:upper 'label' tariff)
    (, 0:19 '-20 yrs'   1.3)
    (,20:29 '20-29 yrs' 1.2)
    (,30:49 '30-49 yrs' 1.0)
    (,50:99 '50- yrs'   0.8)
  )
;
TEXT
 Tariff analysis example.
Endproc
Proc Excel
  listfil(C:\Rapp\Txt\T01.Txt) Xmlfil(C:\Rapp\Xml\T01.Xml)
Endproc
Proc Graf t r f m s u
  listfil(C:\Rapp\Txt\T01.Txt) pdffil(C:\Rapp\Xml\T01.Pdf)
Endproc
```

This program uses the MMT (Marginal Totals) method. A parameter

can be given to use Standard GLM, ie. assuming gamma distributed claim amounts with a constant coefficient of variation.

3. Presenting tariff analysis output

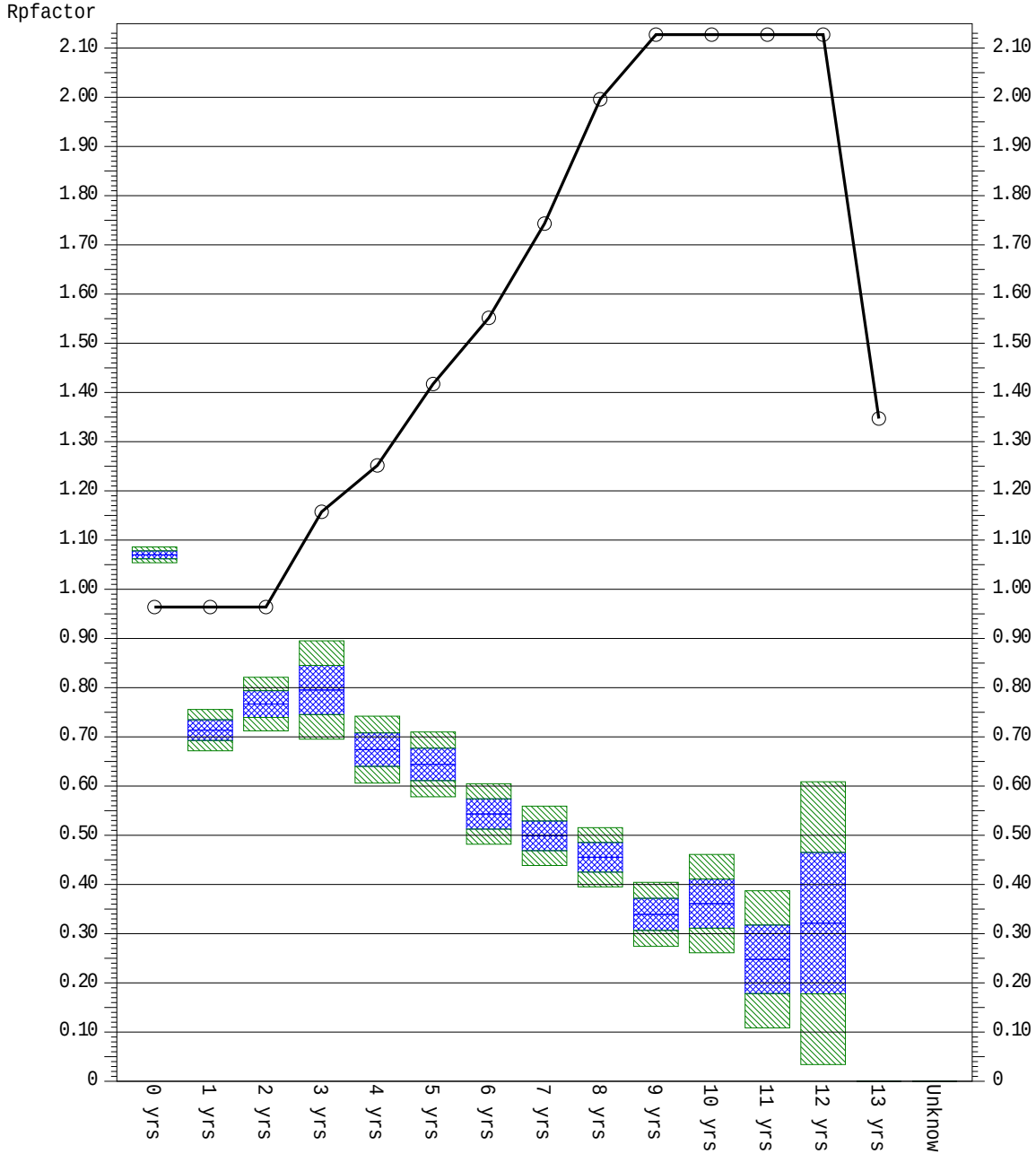
The last two steps of the tariff analysis program produce an Excel table and a graph in Pdf. Start your presentation by showing the graph file.
It is often instructive to make a copy of this file and display two files simultaneously. Showing the factors from the GLM/MMT equations side by side with the univariate factors can give an aha experience. The MMT factors are produced with the parameters t (if tariff factors are available) and r to Proc Graf. The univariate factors are produced with parameter u.

The next two pictures here are graphs over risk premium factors, tariff

factors, and univariate risk premiums. The tariff was set before the actuary (me) was consulted - it reflects the univariate factor ladder. The confidence intervals for risk premium factors are computed with the method MVW, which is described in my paper I first mentioned. There is not necessarily a class with factor 1 and confidence interval length 0 with this method.

Animal age at inception, Parrot insurance

Tariff- and risk premium factors + confidence intervals
Black=tariff, single lines=90% conf, cross lines=60% conf

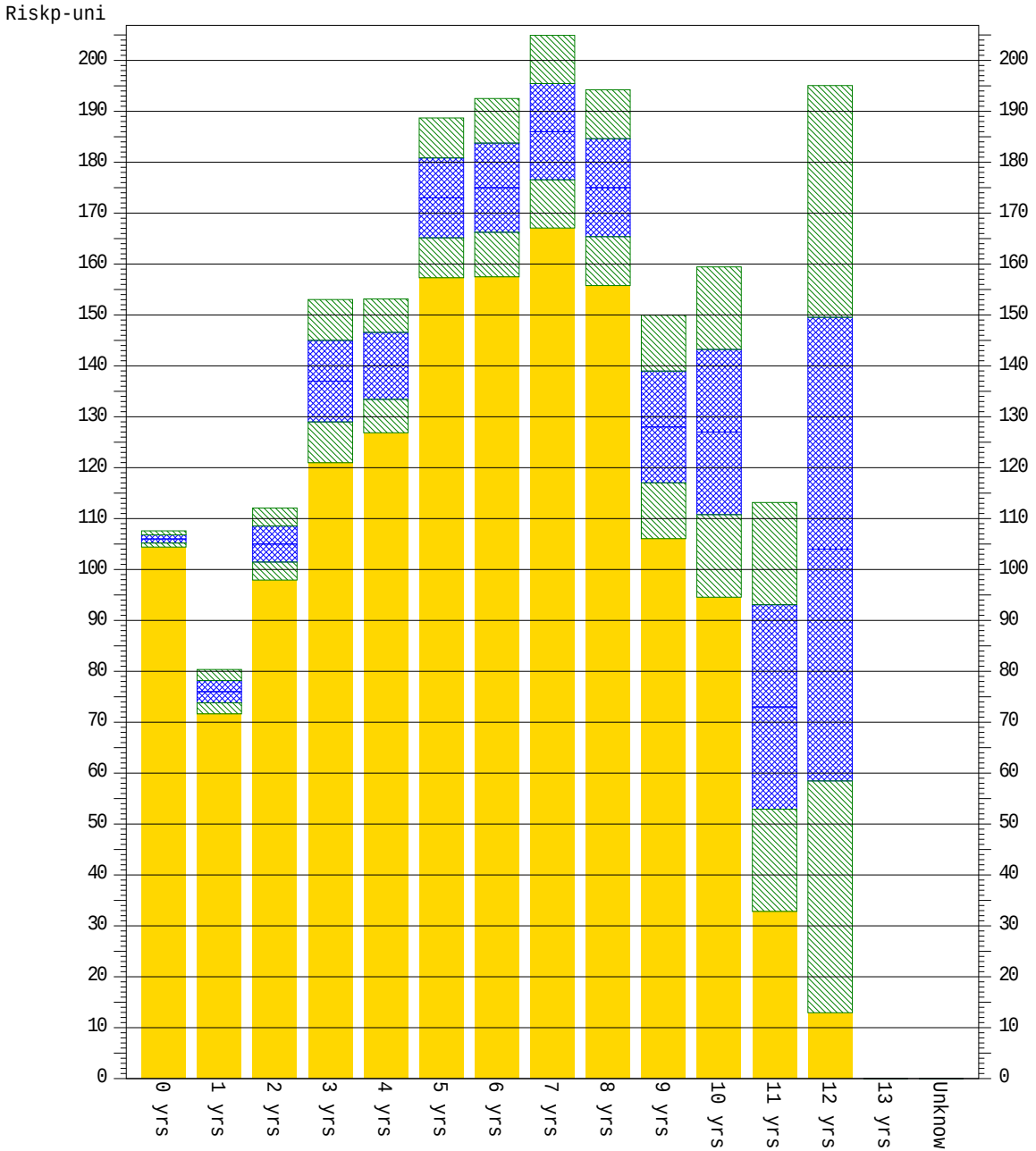


Class:	0 yrs	1 yrs	2 yrs	3 yrs	4 yrs	5 yrs	6 yrs	7 yrs	8 yrs	9 yrs	10 yrs	11 yrs	12 yrs	13 yrs	Unknow
Tarf :	0.964	0.964	0.964	1.157	1.252	1.417	1.552	1.743	1.996	2.127	2.127	2.127	2.127	1.347	0.000
Lo90%:	1.087	0.615	0.566	0.517	0.426	0.367	0.284	0.219	0.170	0.090	0.070	0.021	0.006	0.000	0.000
Point:	1.103	0.654	0.609	0.591	0.473	0.408	0.320	0.248	0.195	0.110	0.096	0.048	0.055	0.000	0.000
Up90%:	1.120	0.692	0.651	0.664	0.520	0.449	0.355	0.278	0.220	0.130	0.121	0.075	0.103	0.000	0.000

Animal age at inception, Parrot insurance

Risk premium univariate with confidence intervals

Single lines=90% confidence, cross lines=60% confidence



Class:	0 yrs	1 yrs	2 yrs	3 yrs	4 yrs	5 yrs	6 yrs	7 yrs	8 yrs	9 yrs	10 yrs	11 yrs	12 yrs	13 yrs	Unknow
Lo90%:	104.409	71.643	97.929	120.973	126.861	157.306	157.489	167.080	155.769	106.054	94.561	32.839	12.940	0.000	0.000
Point:	106.000	76.000	105.000	137.000	140.000	173.000	175.000	186.000	175.000	128.000	127.000	73.000	104.000	0.000	0.000
Up90%:	107.591	80.357	112.071	153.027	153.139	188.694	192.511	204.920	194.231	149.946	159.439	113.161	195.060	0.000	0.000