Robust paradigm applied to parameter reduction in actuarial triangle models

ICASQF 2016

Gary Venter – University of New South Wales

Preliminaries

- Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space = sample space + σ-algebra + probability measure. (To make it clear we are doing measure theory.)
- "Data" is:
  - A mass noun, like "snow." "Snow are white" is true iff snow are white, Tarsky never said.
  - A plural in Latin used as a singular in English, like "agenda." "This meeting doesn't have enough agenda." Cancel it.
  - "Poor data are our biggest problem." Overheard at AIG
- Data is
- IMHO. But then, I don't think "they" can be singular.

Models as Stories We Tell in Math

- "All models are wrong but some are useful"; George Box – maybe Sergio Armani too
- "All models are approximations and some are very good approximations." Andrew Gelman
- Either way: The data was generated by a more complex process than the model specifies.
- A major new viewpoint – Classical and Bayesian statistics assume the data comes from the model process.
- I call the new view the robust paradigm as we want the model to be robust to new data coming from this more complex process. Usual stat tests of fit less reliable here.
- Practical implication is to do out-of-sample testing – everyone is doing that anyway, like in consumer finance, but it is needed under this paradigm.
- Consumer finance asks the question: How can you get a decent return charging 30% interest on credit cards? Answer: securitize 2% mortgages, keep the good stuff.
- Often somewhat simplified models seem to do better out of sample.

Out-of-Sample Testing

- Various forms of hold-out samples – estimate parameters on part of sample and test on the rest
- One popular is rotating 4/5ths: sample divided into 5 portions, each left out in turn. Models compared based on their average performance on the omitted portions – e.g., sum of log-likelihoods.
- Growing in popularity is leave one out (LOO) in which each point is left out one time and the sum of the likelihoods of the missing points compared across models.
- If estimation is fast, this can be automated efficiently enough.

Comparing Fits of Nonlinear Models

- Typical methods like penalized likelihood run into problem of how to count parameters
- Generalized degrees of freedom of Ye fairly good
  - GDOF = sum of derivatives of fitted wrt actual values
  - Agrees with counting for linear and polynomial models
  - How much a data point can pull the model to it makes sense as DOF used up by the point.
  - But typically calculated numerically by reestimating model by tweaking each point, which can add a couple of orders of magnitude to the fitting time
- Still leaves problem that penalized likelihood involves some judgment and is only a view on value of fit.
- LOO is viewed as better than AIC-BIC, and is no more costly than Ye.

Formalized Parameter Reduction

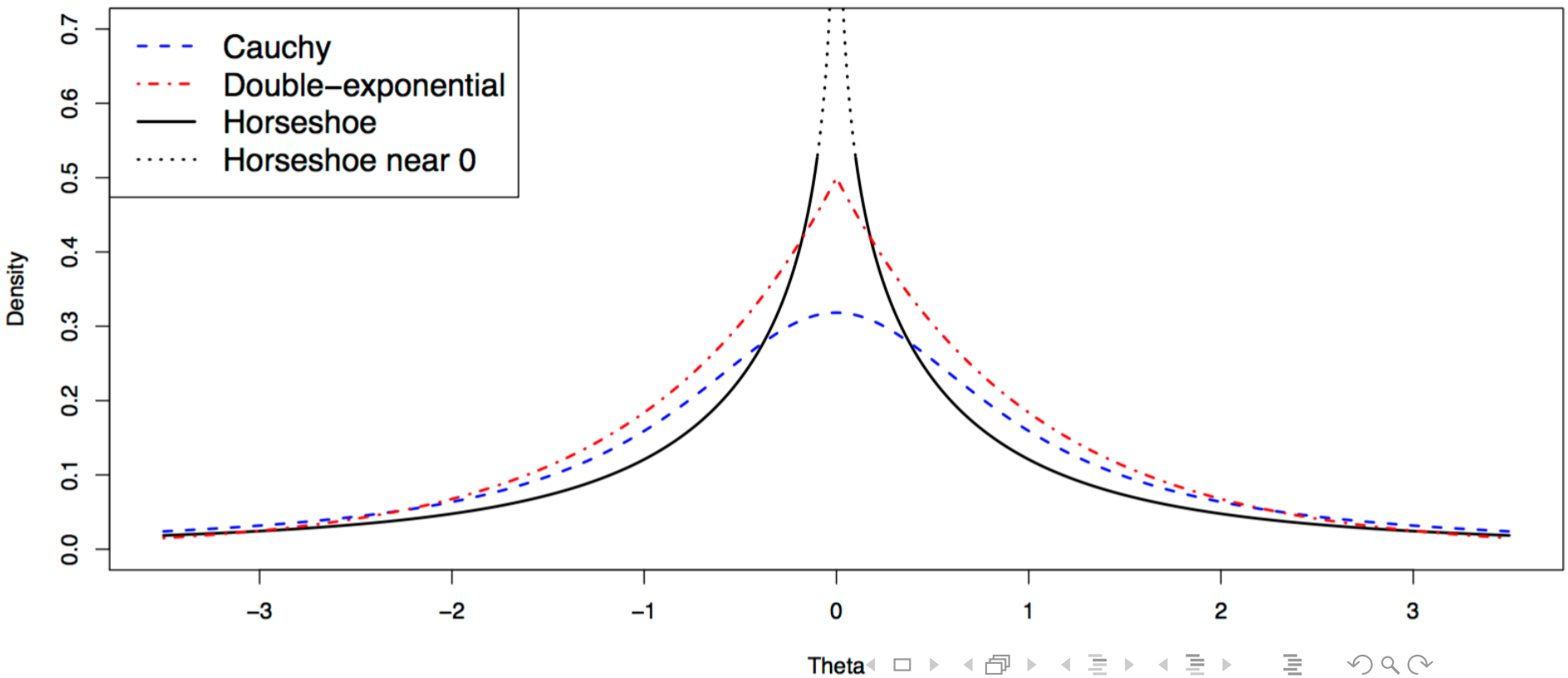
- Used for variable selection and for more robust models
- Two popular methods are LASSO, lately replacing ridge regression, and random effects, used in linear mixed models and generalized linear mixed models – LMM and GLMM.
- LASSO starts by scaling each independent variable to have mean zero and variance one (except the constant).
- Then what is minimized is the NLL plus a selected % of the sum of the absolute values of the coefficients.
- Can start with a lot of variables, and best combination will be kept and coefficients even reduced based on LOO ...
- For LMM some variables are postulated as random effects whose coefficients will be shrunk towards zero unless really needed. E.g., maybe car color in an auto claims regression.
- Each random-effects coefficient b_i is assumed normally distributed with mean zero and variance $d_i\sigma^2$. The d_i's are to be estimated and σ^2 is the regression variance.
- What is maximized is the joint likelihood = likelihood times the normal probability of b: $P(y, b) = P(y|b)P(b)$.

More Details

- Parameter counting tricky for LMM. Tried Le's generalized degrees of freedom in one model. Found that d_i's using up about 60% of dof – more than the regression parameters.
- But could use just one d for all variables. Lasso-like. Or make the variances correlated. Maybe would do that with variables like first letter of last name, month of birth, ...
- I've stopped using LMM in favor of Lasso due to dof issue, but alternatives worth trying.
- Issue with Lasso is selection of shrinkage constant. Loo recommended for that but judgment useful too.
- Advantage of Lasso over ridge regression is parameters actually go to zero.
- Advantage over stepwise in keeping combinations, not one at a time.

Bayesian Version – Shinkage Priors

- Give some parameters priors with mean zero.
  - Lasso-like to have a selected variance d for all those variables, then use loo to find best d but could estimate various variances as in LMM.
  - Double exponential prior pretty common, centered @ 0.
  - Heavier tails than normal, also more weight near zero.
  - Even more so is horseshoe: normal σ^2 mixed by Cauchy.



MCMC

- Estimates posterior of parameters given data by numerical sampling starting with priors for each parameter.
- Yields a distribution of parameters in a numerical simulation.
- Allows estimation of models otherwise intractable.
- But useful for other models too: gets correlated distributions of parameters, loo testing, easy expressions of formulas
- Latest and greatest version is Stan from Columbia.
- Simulates 4+ parallel chains to check convergence.
- Can keep triangle in a rectangle instead of stringing out into a column by making cells you don't want to use large, as in y below:
- for (n in 1:N) { for (u in 1:U) { if (y(n,u)<98) y(n,u)~$normal(m + p[n] + q[u] + r[n + u - 1], sigma_y)$; }}
- Stan compiles models into C so much faster than MCMC interpreted in R directly.

LOO IN MCMC

- With large sample of parameters and likelihood at every point for every sample, we can estimate how parameters would change from leaving out a point by giving more weight to the samples that fit it poorly.
- Been known for 20 years, but gave unstable estimates
- Stan team found improvement: reweighting by "Pareto smoothed importance sampling" – solves stability problem
- Can now do LOO maybe with only a 5% increase in fitting time instead of 20,000%
- Makes out-of-sample testing easy with just a couple of extra lines of code and the loo package in R
- Can run loo on any MCMC output - not just Stan
- Revolutionizes robust fitting by making out-of-sample testing routine.

Selection of Priors

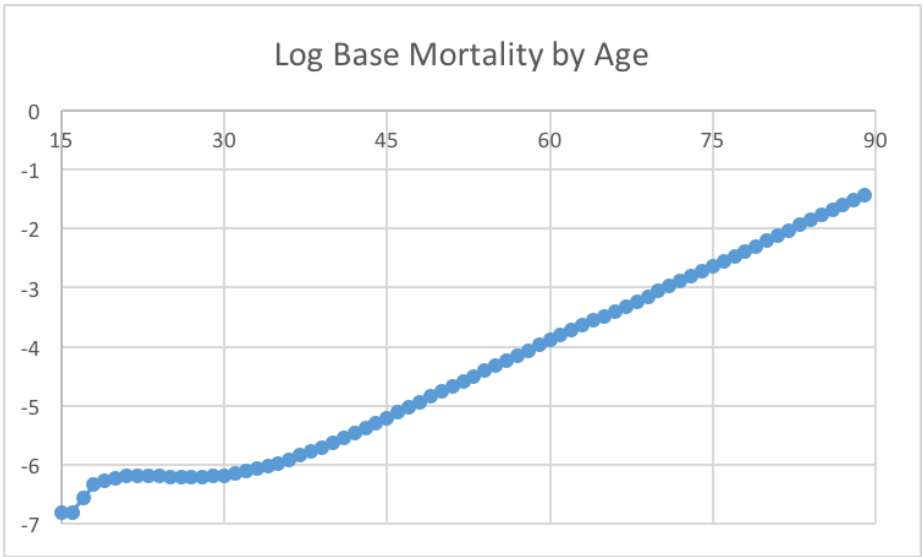
- If you have a good, or strong, belief in where the parameters should be, use that as a prior
- Non-informative priors useful if you pick the right ones.
- For a real number that could be positive or negative uniform prior works well
- Typically gives posterior variance close to classical estimation variance.
- But patently absurd as a belief: Parameter is probably very large in absolute value, but don't know sign. Really?
- Probably implemented on a finite range anyway based on machine precision.
- For positive parameter log uniform works well. Uniform prior on positive reals has infinite pull up but not down and may bias posterior upward.
- An informative absurd prior is gamma(0.001, 0.001). Mean is 1, but median is 10^{-27}.

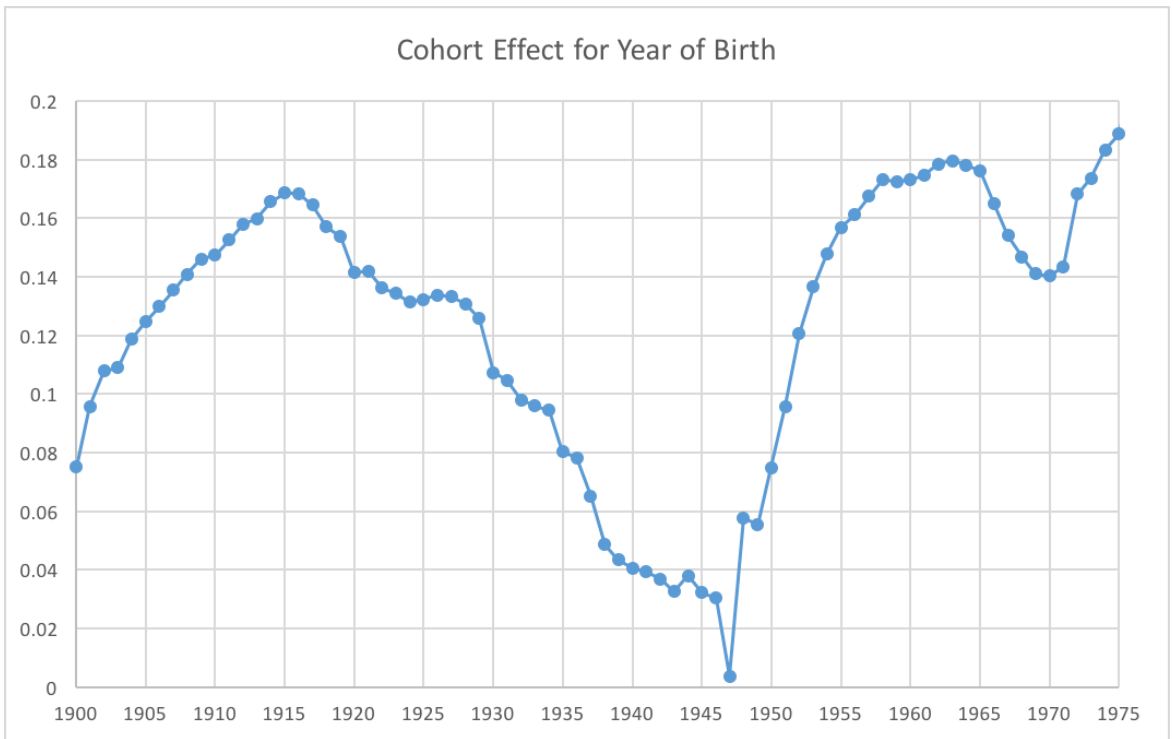
Three Trends Model

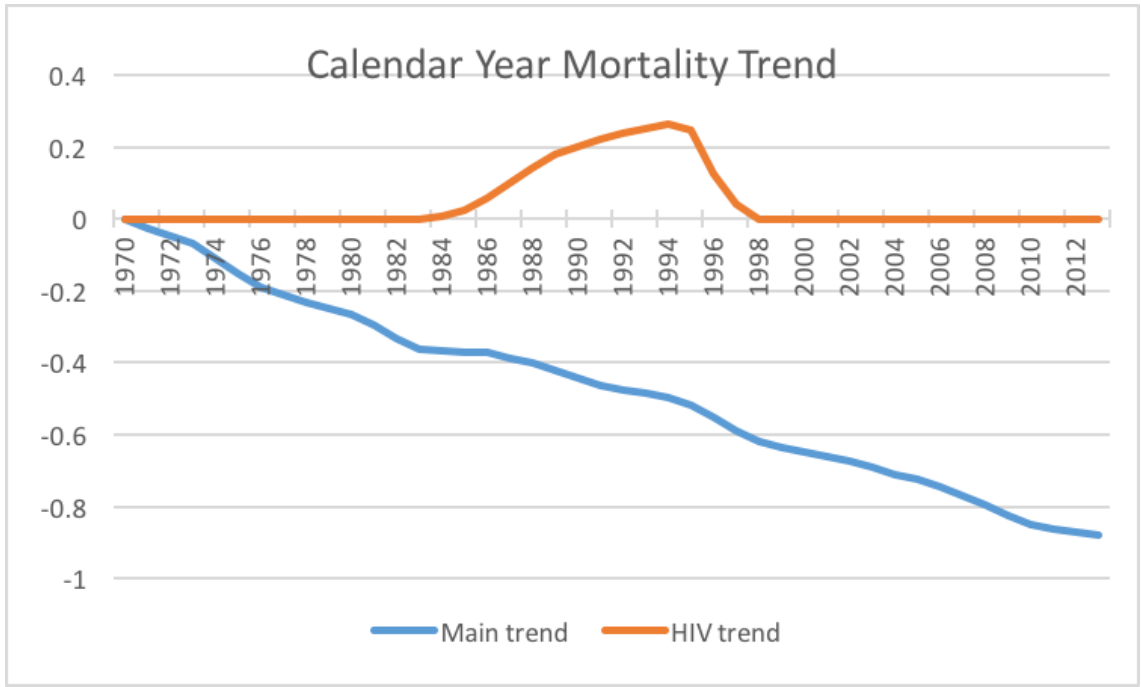
- Fairly usual approach, sort of started by Taylor 1977.
  - y(n,u) = log of claims for year n and delay u
  - For reserves, mean modeled as: p(n) + q(u) + r(n + u)
  - Mortality model has interactions: p(n)x(u) + q(u) + r(n + u)y(u)
  - Too many parameters. Casualty actuaries make piecewise linear. Life actuaries smooth parameters with cubic splines.
- Piecewise linear means changes in levels = trends are piecewise constant, so modeling focus is on 2nd differences = trend changes, with a lot of them zero.
- E.g., Insureware has a wizard to tell them which ones to make non-zero then software estimates those.
- I have been trying parameter shrinkage methods on the trend changes like random effects, Lasso – classical and Bayesian. Shrink 2nd differences -> 0, then add up to get trends then levels.

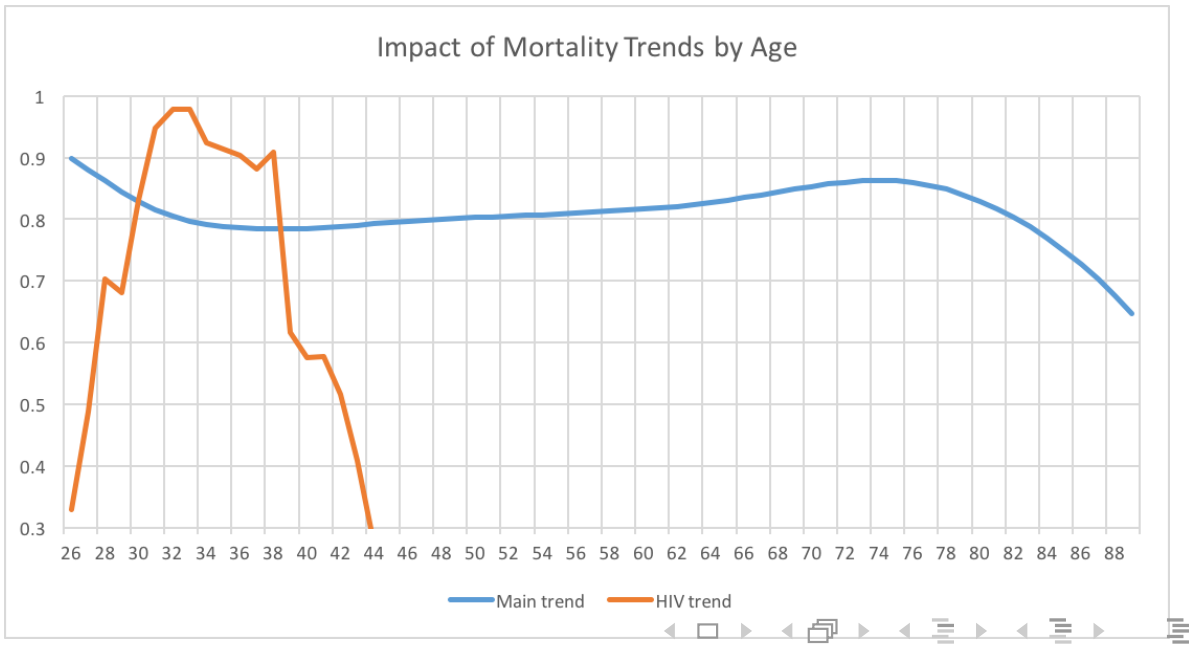
Mortality Trends

- Mortality data looks like loss triangles, with rows by year of birth n, columns age at death u, diagonals year of death n+u. Log of fraction of those alive who died during the year is the data y
- Tried model with interactions, based on Renshaw - Haberman model, y(n,u) = p(n) + q(u) + r(n+u)z(u)
- The base mortality by age is q; r is the time trend by diagonal, which impacts more strongly at some ages, reflected by z(u); p is the year-of-birth, or cohort, effect.
- Used Human Mortality Database for US males, ages 16-99 for calendar years 1971-2010, and cohorts 1881-1955.
- Parameters given shrinkage priors were trend changes, or 2nd differences in p, q, r, and z. Used loo to determine shrinkage level, i.e., variance of double exponential prior.
- Time trend actually has different age impacts in different years. Trend leveled out 1985-1995 so tried an additional trend then with its own age impact. This added term s(n+u)w(u) to mean. Need shrinkage to do all this.
- High degree of smoothing given by loo.


Log Base Mortality by Age
0
-1
-2
-3
-4
-5
-6
-7
15
30
45
60
75
90


Cohort Effect for Year of Birth
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975


Calendar Year Mortality Trend
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012
Main trend
HIV trend


Impact of Mortality Trends by Age
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88
Main trend
HIV trend

Bayesian vs. Classical Lasso

- Model with interactions y(n,u) = p(n) + q(u) + r(n+u)z(u)
- Typical classical estimation starts with initial values for r or z, estimating contingent on that, freezing those, estimating the first one, back and forth until they converge.
  - Advantages of Bayesian include direct estimation without this iteration, not needing to put into a single column, getting distributions of parameters including correlations, and being able to easily get best shrinkage with loo.
- Problem with this model, especially with interactions, is parameters can offset each other giving local maxs, even with enough constraints for identifiability. Some maxs are interpretable as meaningful trends, some are not.
  - Classical estimation can use good starting values and stay in that vicinity. Bayesian can get a lot of parameter sets that look strange even if they fit. Even if all simulated sets of parameters are good fits, average might not be – then hard to say what the estimated parameters actually are.
  - Fitting's actual result is not the parameters but the distribution of predictive outcomes, but averages of one or two chains usually can represent parameter fits.

Workers Comp Loss Triangle Example

- Taylor & McGuire use this triangle to tell us everything we need to know about GLM in reserving

		0	1	2	3	4	5	6	7	8	9
1988	1	41,821	34,729	20,147	15,965	11,285	5,924	4,775	3,742	3,435	2,958
1989	2	48,167	39,495	24,444	18,178	10,840	7,379	5,683	4,758	3,959	
1990	3	52,058	47,459	27,359	17,916	11,448	8,846	5,869	5,391		
1991	4	57,251	49,510	27,036	20,871	14,304	10,552	7,742			
1992	5	59,213	54,129	29,566	22,484	14,114	10,000				
1993	6	59,475	52,076	26,836	22,332	14,756					
1994	7	65,607	44,648	27,062	22,655						
1995	8	56,748	39,315	26,748							
1996	9	52,212	40,030								
1997	10	43,962									

- Start with exploratory analysis: Two quick looks at the data.
  - Calculate cumulative-to-incremental developments factors, then take ratios of individual factors to column averages, subtract 1, color code big/little; positive/negative and rotate so rows are calendar years
  - Develop losses to ultimate then take ratios of paid in column to ultimate for each row.

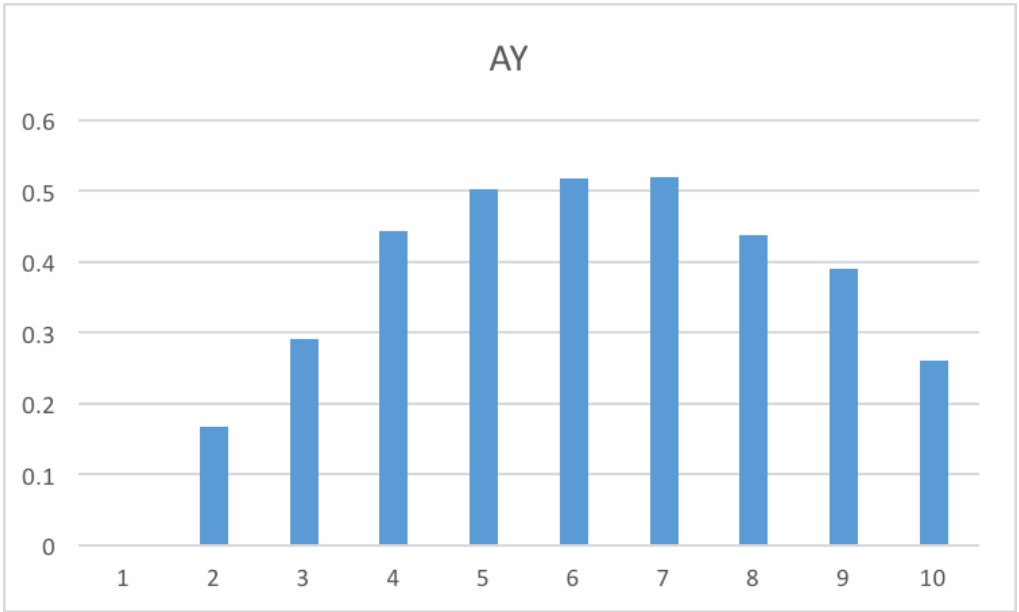
	1	2	3	4	5	6	7	8
1	2%							
2	1%	1%						
3	12%	7%	4%					
4	6%	5%	3%	13%				
5	12%	-3%	-11%	-6%	-14%			
6	7%	-0.03%	-1%	-11%	-6%	-5%		
7	-16%	-8%	-0.48%	5%	2%	-1%	-8%	
8	-15%	-6%	2%	-3%	13%	-8%	2%	-0.2%
9	-6%	7%	4%	4%	0.4%	12%	4%	0.2%

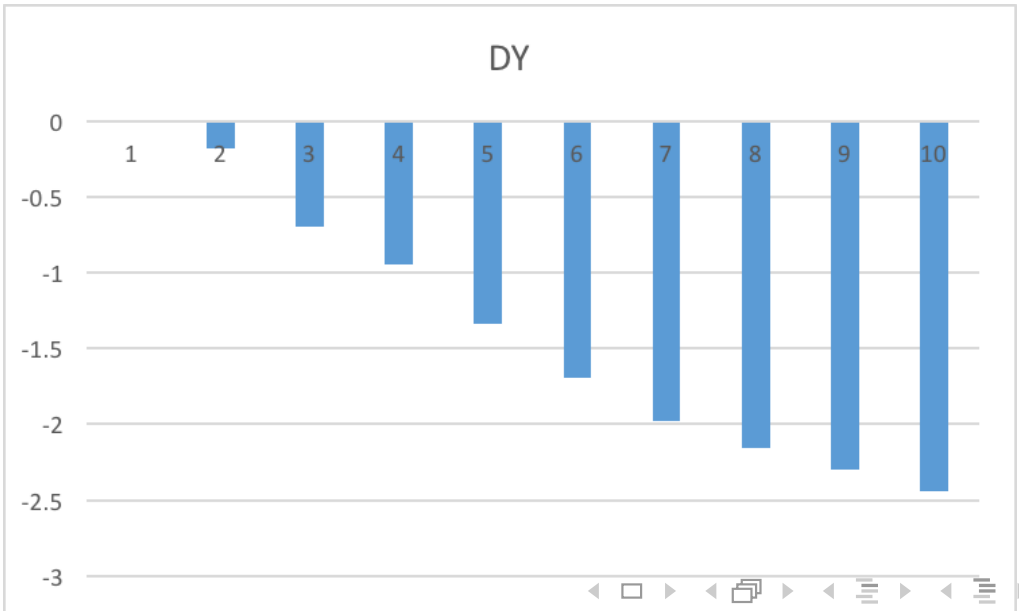
- Factor ratios. 1^{st} 4 rows positive, next 4 mostly negative
- Suggests some change in calendar year trend took place

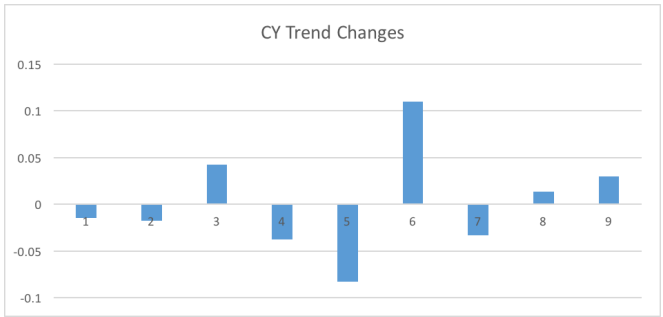
	0	1	2	3	4	5	6	7	8
1	29%	24%	14%	11%	8%	4%	3%	3%	2%
2	29%	24%	15%	11%	7%	4%	3%	3%	2%
3	28%	26%	15%	10%	6%	5%	3%	3%	
4	28%	25%	13%	10%	7%	5%	4%		
5	28%	26%	14%	11%	7%	5%			
6	29%	25%	13%	11%	7%				
7	32%	22%	13%	11%					
8	31%	21%	15%						
9	30%	23%							

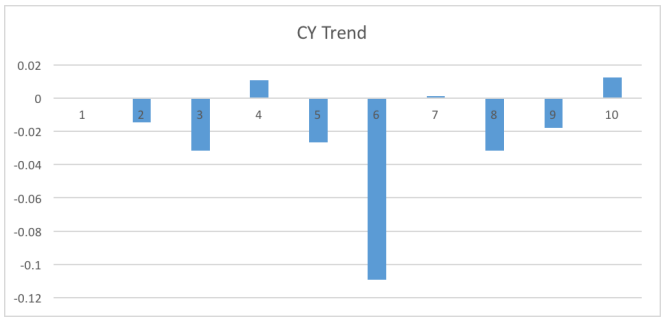
- Payout pattern. Shift from lag 1 to lag 0 last 3 – 4 years
- But complicated by calendar-year shift, so we will fit that first and see if payout shift remains.

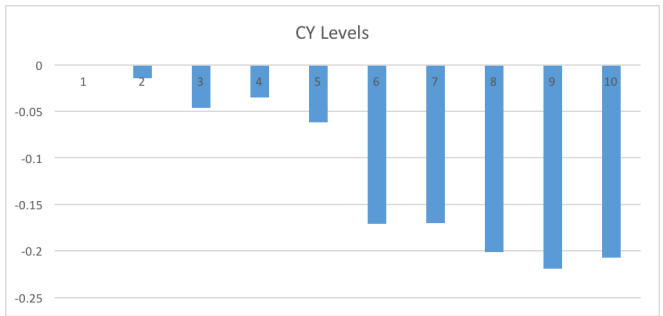
AY & DY Levels Look Pretty Smooth





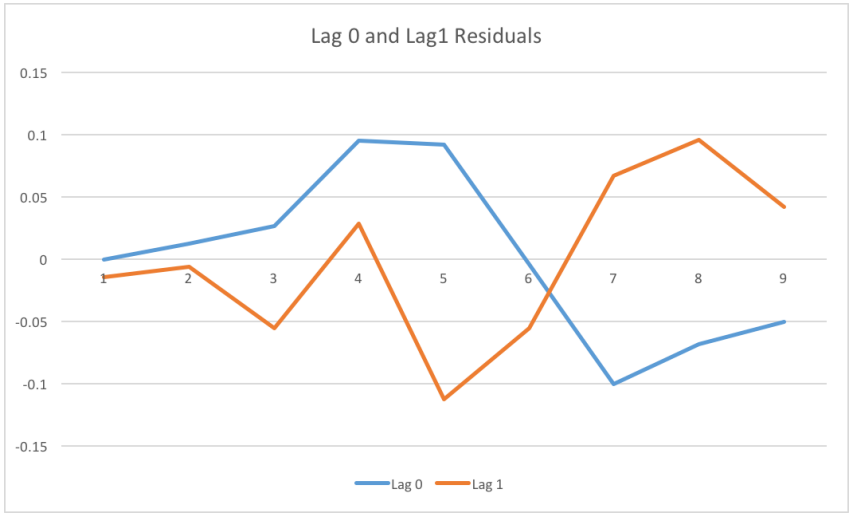






- CY trend shows big jump down in 1993. AY driven by wage inflation, CY by medical – wage. Single jump could be a reform, but wearing off by end when CY trend again positive.

	0	1	2	3	4	5	6	7	8	9	
1	-0.0001924	-0.0142322	-0.0102601	-0.0150715	-0.0933068	0.08519159	0.01923436	0.05630786	-0.0194738	0.00105833	0.00925543
2	0.01261907	-0.0060851	-0.0242343	-0.0029425	0.00624825	0.03521971	-0.0180514	-0.0331258	0.01932401		-0.011028
3	0.02662277	-0.0554801	-0.0400109	0.02584092	0.07627243	-0.0543661	0.05545896	-0.0223151			0.01202282
4	0.09506255	0.02832493	0.01536021	0.02700008	-0.0254809	-0.0957714	-0.0565732				-0.012077
5	0.09228061	-0.1125751	-0.0154694	-0.0216719	0.02763631	0.02770303					-0.002096
6	-0.0042547	-0.0556989	0.0667669	-0.0155591	0.01248208						0.003736
7	-0.1006665	0.06705556	0.04121433	-0.0170865							-0.009483
8	-0.0682318	0.09560563	-0.0157693								0.01160458
9	-0.0505447	0.04196046									-0.008584
10	0.00415631										0.00415631
	0.00685121	-0.0111248	0.01759748	-0.0194906	0.00385138	-0.0020232	6.8726E-05	0.00086692	-0.0001498	0.00105833	-0.002494



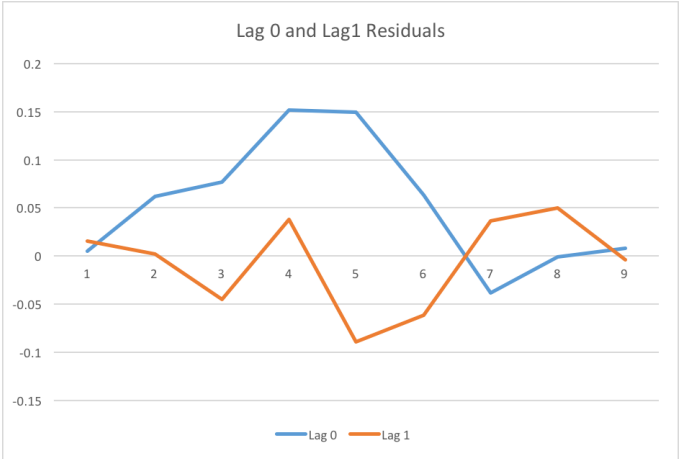
- Some but not much shrinkage in this fit. Small residuals by row and column but all would be 0 if no shrinkage.
- Payout pattern shift still there at bottom of cols 0, 1.

Try Interaction Terms for Payout Shifts

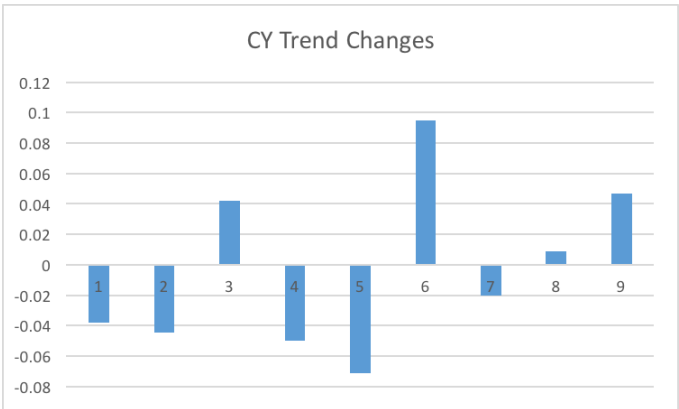
- Meyers has a number of models with trends for payout shifts. Here combining with CY changes as well.
- Try interaction of DY and AY: p(n) + w(n)x(u) + q(u) + r(n + u)
- Started trend changes for w at last row, for x in last column.
- Gave bigger impact in earlier columns.
- Zeros at end of first column are because these are the base cells for identifiability. They have the largest DY values.

DY term after interaction									
-0.0846236	-0.2222793	-0.6875312	-0.8748535	-1.1997081	-1.594766	-1.7877335	-1.905103	-1.9601988	-2.0607415
-0.0547095	-0.2309704	-0.6827897	-0.8662319	-1.2002981	-1.5819962	-1.7773879	-1.8915242	-1.9601988	
-0.0562694	-0.2305172	-0.6830369	-0.8666815	-1.2002674	-1.5826621	-1.7779274	-1.8922323		
-0.0602586	-0.2293582	-0.6836692	-0.8678312	-1.2001887	-1.584365	-1.779307			
-0.0616277	-0.2289604	-0.6838862	-0.8682258	-1.2001617	-1.5849495				
-0.0420673	-0.2346434	-0.6807858	-0.8625882	-1.2005475					
-0.0177594	-0.2417056	-0.6769329	-0.8555824						
-0.0059299	-0.2451425	-0.6750579							
0	-0.2468654								
0									

Interaction Does Fit Shift

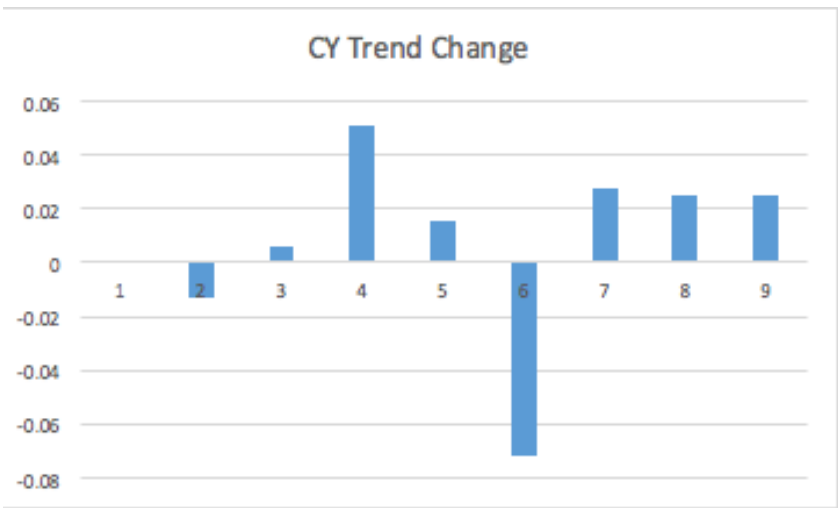


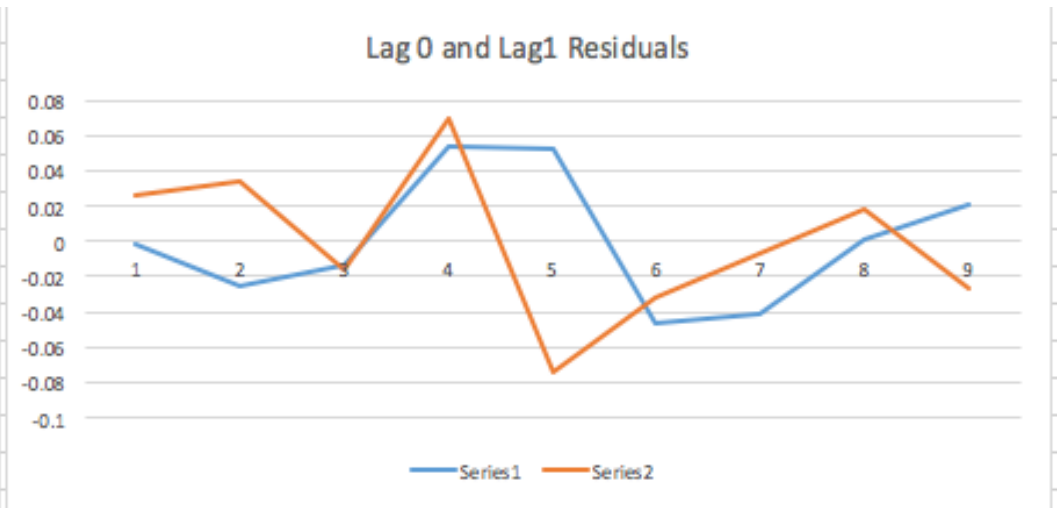
- But column 5 residuals are bigger. CY trend now increasing a bit more at end.
- Residual std dev drops to 3.7% from 5.2% so means better represent data.



Alternatives to Interaction

- Can put any DY trending model in there like one of Meyers' – with functions applied to row and column numbers – and still keep CY trends.
- Tried simple one parameter added to rows 7 – 10 in first column and subtracted in 2nd.
- Fit a little worse than interaction with residual std dev of 4.2% – but fit AY 5, DY 0 and 1 better. Even more CY trend change at end





Summary

- Can do parameter shrinkage in classical or Bayesian models
- Select degree of shrinkage by judgment or out-of-sample testing
- Out-of-sample testing best way to evaluate models known to be oversimplified, and easiest to do with loo package for MCMC.
- Multiple trend models provide intuitive stories of what is happening in triangle data, but complicated versions can have problems with local max. Selecting among MCMC chains seem to be a practical way to deal with this.