“LDF Curve Fitting”

Dave Clark’s, 2003 *Forum* paper

An R Schematic*

Daniel Murphy, FCAS, MAAA

*R code shown below in `Courier` font

Clark's Longitudinal Reserving Model

- Dave Clark's 2003 Forum paper describes a *Longitudinal* model for reserving

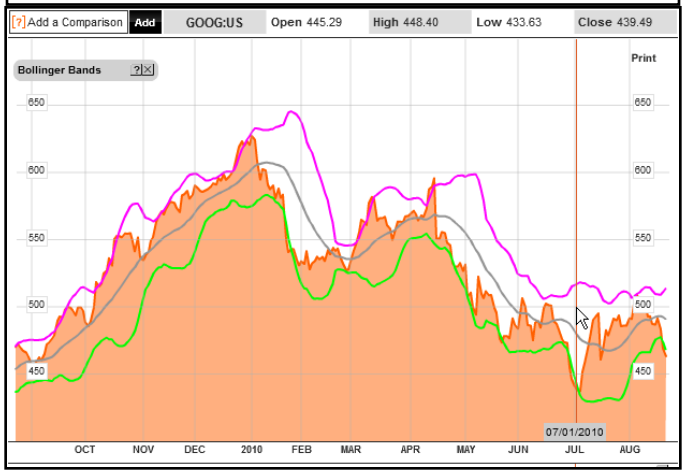
- More familiar term: *Time Series Analysis*
  - Single subject observed over many time periods
  - Example: daily closing price of a stock

- Less familiar term: *Longitudinal Analysis*
  - Many subjects observed over few time periods
  - Example: five-year survey of group of students

Examples

Time Series Data

google daily stock price

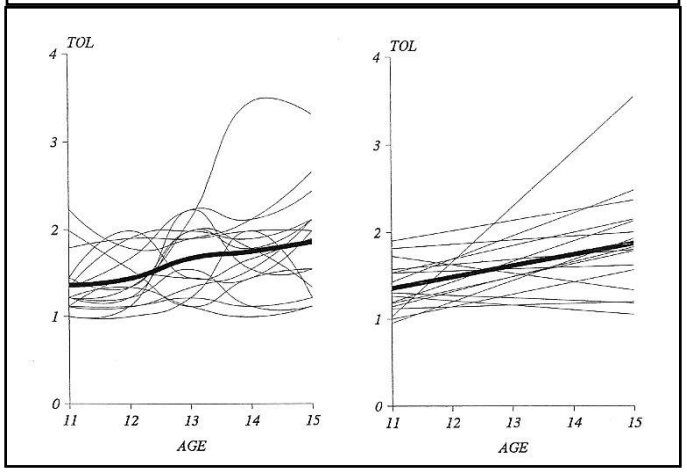


- ~250 observations of one subject
- Bollinger bands show ±2 standard deviations
  - Can trigger buy/sell decisions

www.bloomberg.com

Longitudinal Data

Teenage tolerance of "deviant behavior"



- Five observations of ~1600 subjects
- Adolescents tend to become more tolerant with age

Individual and average (bold) growth pattern fits: nonparametric smoother on left, ordinary least squares (OLS) on right

National Youth Survey, Raudenbush & Chan, 1992

"Spaghetti plots" from Singer & Willett, *Applied Longitudinal Data Analysis*, Oxford, 2003

Actuarial Example of Longitudinal Dataset

Clark looks at well-researched *Taylor-Ashe* data

Paid Losses by age (months)										
AY	12	24	36	48	60	72	84	96	108	120
2001	357,848	1,124,788	1,735,330	2,218,270	2,745,596	3,319,994	3,466,336	3,606,286	3,833,515	3,901,463
2002	352,118	1,236,139	2,170,033	3,353,322	3,799,067	4,120,063	4,647,867	4,914,039	5,339,085	
2003	290,507	1,292,306	2,218,525	3,235,179	3,985,995	4,132,918	4,628,910	4,909,315		
2004	310,608	1,418,858	2,195,047	3,757,447	4,029,929	4,381,982	4,588,268			
2005	443,160	1,136,350	2,128,333	2,897,821	3,402,672	3,873,311				
2006	396,132	1,333,217	2,180,715	2,985,752	3,691,712					
2007	440,832	1,288,463	2,419,861	3,483,130						
2008	359,480	1,421,128	2,864,498							
2009	376,686	1,363,294								
2010	344,014									

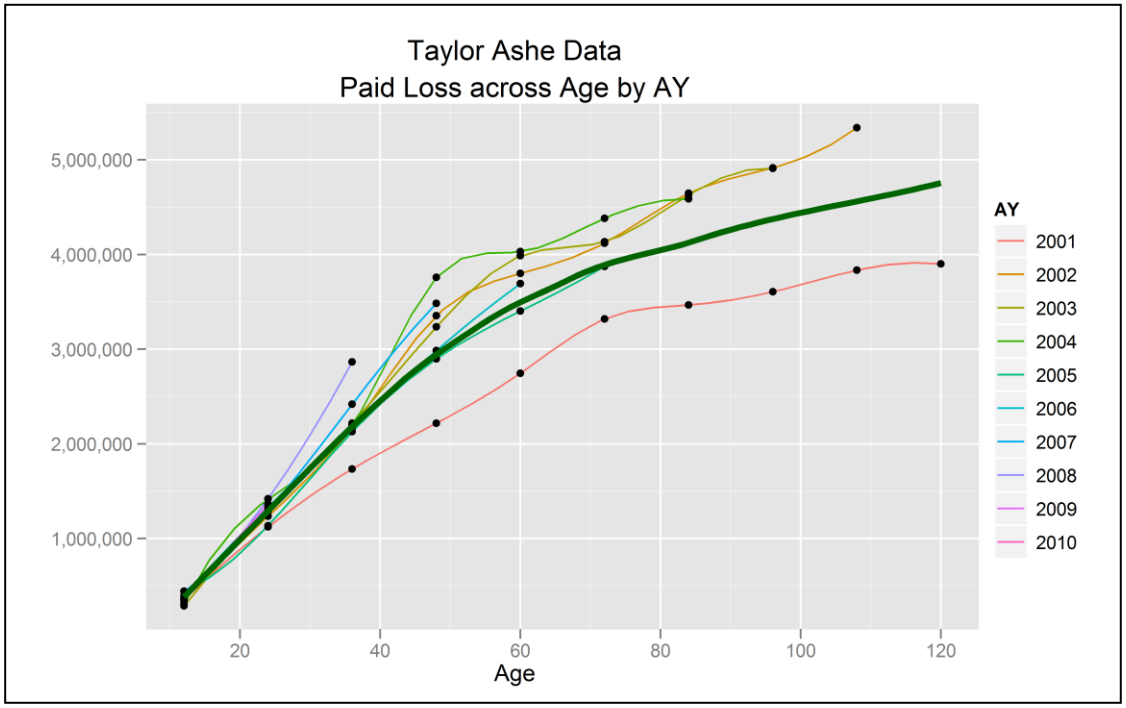
updated AY labels

AY	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120
2001	3.143	1.543	1.278	1.238	1.209	1.044	1.040	1.063	1.018
2002	3.511	1.755	1.545	1.133	1.084	1.128	1.057	1.086	
2003	4.448	1.717	1.458	1.232	1.037	1.120	1.061		
2004	4.568	1.547	1.712	1.073	1.087	1.047			
2005	2.564	1.873	1.362	1.174	1.138				
2006	3.366	1.636	1.369	1.236					
2007	2.923	1.878	1.439						
2008	3.953	2.016							
2009	3.619								

- In longitudinal terms
  - 10 “subjects” (accident years)
  - One to ten observations of each subject
- Growth patterns
  - Traditionally, actuaries select loss development (growth) patterns by
    - Scrutinizing loss and link ratio triangles
    - Exercising actuarial judgment regarding years to include/exclude, types of averages to use, …
  - However, graphical techniques can reveal unanticipated patterns

AY 2001 "Looks Different" in Spaghetti Plot

Important to seek business explanation for observed anomalies



- Stricter underwriting in first year of program?
- Consider omitting that year's data if prudent

Individual and average (bold) nonparametric growth pattern fits
– See also R. E. Sherman, "Extrapolating, Smoothing and Interpolating Loss Development Factors," *PCAS*, 1984
ggplot2 graphics package for R by Hadley Wickham, Rice University

"Abandon your triangles!" – *Dave Clark, 2003 Forum*

Longitudinal data easier to analyze in *long format*

AY	12	24	36	48	60	72	84	96		108	120
2001	357,848	1,124,788	1,735,330	2,218,270	2,745,596	3,319,994	3,466,336	3,606,286	…	3,833,515	3,901,463
2002	352,118	1,236,139	2,170,033	3,353,322	3,799,067	4,120,063	4,647,867	4,914,039	…	5,339,085	
2003	290,507	1,292,306	2,218,525	3,235,179	3,985,995	4,132,918	4,628,910	4,909,315			
2004	310,608	1,418,858	2,195,047	3,757,447	4,029,929	4,381,982	4,588,268				
2005	443,160	1,136,350	2,128,333	2,897,821	3,402,672	3,873,311					
2006	396,132	1,333,217	2,180,715	2,985,752	3,691,712						
2007	440,832	1,288,463	2,419,861	3,483,130							
2008	359,480	1,421,128	2,864,498								
2009	376,686	1,363,294									
2010	344,014										

Wide format

AY	Age	Paid Loss
2001	12	357,848
2002	12	352,118
2003	12	290,507
2004	12	310,608
2005	12	443,160
2006	12	396,132
2007	12	440,832
2008	12	359,480
2009	12	376,686
2010	12	344,014
2001	24	1,124,788
2002	24	1,236,139
2003	24	1,292,306
2004	24	1,418,858
2005	24	1,136,350
⋮	⋮	⋮
2001	108	3,833,515
2002	108	5,339,085
2001	120	3,901,463

Long format

- *Wide format* (matrix): functional for few subjects
  - Can see all observations of all subjects
- *Long format* (table): better for many subjects
  - Statistical routines (e.g., lm, glm, nle) prefer tables ("`data.frames`" in R)
  - Need graphical techniques to see all observations of all subjects
- R code to convert from matrix to table

```
reshape(as.data.frame(M), direction="long", varying=colnames(M), times=colnames(M),
        ids=rownames(M), v.names="Paid Loss", idvar="AY", timevar="Age")
```

or

```
library(reshape)
melt(M)
```

("reshape" package by Hadley Wickham)

or

```
library(ChainLadder)
as.data.frame(as.triangle(M))
```

("ChainLadder" package by Markus Gesmann)

Clark Models Incremental Loss in Long Format

Utilizes maximum likelihood technique

Clark's Table 1.1
Incremental Loss by Development Period

i	AY	Age	Age.from	Age.to	Incr_Loss
1	2001	12	0	12	357,848
2	2001	24	12	24	766,940
3	2001	36	24	36	610,542
4	2001	48	36	48	482,940
5	2001	60	48	60	527,326
6	2001	72	60	72	574,398
7	2001	84	72	84	146,342
8	2001	96	84	96	139,950
9	2001	108	96	108	227,229
10	2001	120	108	120	67,948
11	2002	12	0	12	352,118
12	2002	24	12	24	884,021
13	2002	36	24	36	933,894
⋮	⋮	⋮	⋮	⋮	⋮
46	2007	12	0	12	440,832
47	2007	24	12	24	847,631
48	2007	36	24	36	1,131,398
49	2007	48	36	48	1,063,269
50	2008	12	0	12	359,480
51	2008	24	12	24	1,061,648
52	2008	36	24	36	1,443,370
53	2009	12	0	12	376,686
54	2009	24	12	24	986,608
55	2010	12	0	12	344,014

- Assume
  - Losses follow loglogistic (inverse power curve) growth pattern
    - $G(age, \omega, \theta) = \dfrac{1}{1 + \left(\dfrac{\theta}{age}\right)^{\omega}}$
    - Two parameters (ω, θ) to estimate
  - Each incremental loss in "long triangle" is an independent observation from same ODP/growth-curve process
  - Each AY has own ultimate level
    - "Model #2: LDF Method" per Clark
    - Ten additional parameters (U, a vector) to estimate
- Maximum likelihood technique
  - Find parameters (U, ω, θ) that maximize chance that actual incremental losses will be the values observed

Maximum Likelihood Estimation (MLE) with R

Schematic using `optim` function

- loglikelihood function to be maximized (Clark, p. 51)

$$L = \sum_i Incr\_Loss_i \times \ln(Expected\_Incr\_Loss_i) - Expected\_Incr\_Loss_i$$

- Function definitions in R
  - Growth function

```
G = function(age, ω, θ) 1/(1+(θ/age)^ω)
```

"%-of-ultimate G is a function of `age`, ω and θ"

"and here is the formula"

  - Expected incremental loss function

```
E_Incr_Loss = function(U, Age.from, Age.to, ω, θ)
      U * (G(Age.to, ω, θ) - G(Age.from, ω, θ))
```

  - Finally, loglikelihood function

```
L = function(Incr_Loss, Age.from, Age.to, U, ω, θ)
     sum(Incr_Loss*log(E_Incr_Loss(U, Age.from, Age.to, ω, θ))
          - E_Incr_Loss(U, Age.from, Age.to, ω, θ))
```

vector arithmetic → one-line formulas

- Ask `optim` to find parameters (U, ω, θ) that maximize `L`
  - Parameter starting values: U = current (cumulative) diagonal, ω = 3, θ = median age of observations
  - `method="L-BFGS-B"` (`optim` has many search methods from which to choose)
  - `hessian=TRUE` (discussed below)

And the Parameter Search Winners Are …

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Values from
U =	**5,044,133**	**7,166,499**	**6,923,028**	**6,904,448**	**6,357,725**	**6,844,465**	**7,763,198**	**8,543,271**	**6,996,314**	**7,184,376**	**R**
	5,050,867	*7,184,079*	*6,939,399*	*6,917,862*	*6,372,348*	*6,867,980*	*7,780,515*	*8,590,793*	*7,033,659*	*7,261,205*	*paper*
	omega =	**1.436**	*1.434294*								
	theta =	**48.418**	*48.6249*		σ^2	**64,422**	*65,029*				

- Estimates reflect a parameterization refinement
  - Clark suggests that a more appropriate measure of accident year *age* is length of time from evaluation date to average date of loss (6/30/AY) rather than to beginning of AY
  - Subtract 6 mos. from Table 1.1 *age* columns before calling `optim`
- σ^2 = average ratio of squared residuals to expected loss
  - Technically a model parameter, but assumed known and equal to sample estimate
  - `sigma2 = sum((Incr_Loss-E_Incr_Loss)^2/E_Incr_Loss)/(55-12)`
- Model is parameterized, can now estimate unpaid claim liabilities
  - Point estimates
  - Risk estimates that reflect parameter uncertainty

Unpaid Claim Liability Estimates by AY

Central estimates, standard errors

AY	PaidLosses	Est'dReserves	ProcessSE	ProcessCV	ParameterSE	ParameterCV	TotalSE	TotalCV
2001	3,901,463	665,308	207,028	31.1%	156,187	23.5%	259,336	39.0%
2002	5,339,085	1,155,322	272,816	23.6%	253,807	22.0%	372,621	32.3%
2003	4,909,315	1,362,639	296,284	21.7%	294,841	21.6%	417,989	30.7%
2004	4,588,268	1,662,754	327,289	19.7%	352,671	21.2%	481,140	28.9%
2005	3,873,311	1,883,214	348,311	18.5%	396,829	21.1%	528,009	28.0%
2006	3,691,712	2,510,629	402,169	16.0%	511,909	20.4%	650,993	25.9%
2007	3,483,130	3,540,981	477,617	13.5%	697,696	19.7%	845,516	23.9%
2008	2,864,498	4,888,860	561,205	11.5%	956,264	19.6%	1,108,779	22.7%
2009	1,363,294	4,984,450	566,665	11.4%	1,214,601	24.4%	1,340,285	26.9%
2010	344,014	6,216,458	632,833	10.2%	2,795,794	45.0%	2,866,521	46.1%
Total	34,358,090	28,870,615	1,363,784	4.7%	4,620,541	16.0%	4,817,603	16.7%

See Clark table, p. 65

By comparison, 13% with Mack Method (Dr. T. Mack, *ASTIN*, 1993)

- Est'dReserves* (R) = $U \cdot (1 - G(age_{ay}, \omega=1.436, \theta=48.4))$ — equivalent to chain ladder
- $ProcessSE^2 = Est'dReserves \cdot \sigma^2$ — thanks to ODP assumption
- "Information matrix" + "delta method" → $ParameterSE^2$ — see next page
- $TotalSE^2 = ProcessSE^2 + ParameterSE^2$
- *Caution! Unwise to extrapolate growth curve blindly*
  - Clark truncates fat-tailed inverse power curve at 20 years
    - 234 months from average date of loss
  - Revised formula: $Est'dReserves = U \cdot (G(234, \omega, \theta) - G(age_{ay}, \omega, \theta))$

* In 2003 parlance, "reserves" = "estimates of unpaid claim liabilities"

Parameter Risk Formulas Are Complicated

Combination of calculus, linear algebra, advanced statistics

Hessian

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	omega	theta
2001	-0.00000015										-0.14990	0.00519
2002		-0.00000010									-0.14170	0.00564
2003			-0.00000010								-0.12792	0.00612
2004				-0.000000096							-0.10626	0.00661
2005					-0.000000096						-0.07374	0.00706
2006						-0.000000079					-0.02711	0.00737
2007							-0.000000058				0.03518	0.00734
2008								-0.000000039			0.10658	0.00660
2009									-0.000000028		0.15504	0.00465
2010										-0.000000007	0.09446	0.00134
2011	-0.14990	-0.14170	-0.12792	-0.10626	-0.07374	-0.02711	0.03518	0.10658	0.15504	0.09446	-20,853,650	-233,499
2012	0.00519	0.00564	0.00612	0.00661	0.00706	0.00737	0.00734	0.00660	0.00465	0.00134	-233,499	-9,466

- Hessian (*H*): matrix of 2nd partial derivatives of loglikelihood function
  - Calculus formulas in Appendix A of paper → H above
  - Note: numerical method version of hessian is available from `optim`
    - convenient • avoids tedious calculus • unavoidable for complex models
- Fisher Information matrix (I) ≡ negative of expected value of *H*: I = -E(*H*)
  - Clark estimates I by $-\sigma^2 H$
- Rao-Cramér theorem: covariance matrix Σ of parameter estimates is bounded below by inverse of I (Hogg & Craig, 4th ed., chap. 11)
  - Clark approximates Σ by I^{-1}
- Delta method: linear approximation of variance of a function of MLE estimates
  - Function of interest in our case is Est'dReserves (R)
  - pre-, post-multiply Σ by gradient of R

continued next page

Parameter Risk (cont.)
R schematic for delta method

- R code: `-sigma2 * dR %*% solve(H,dR)`
- Gradient `dR`: vector of length 12 (Clark p. 55)

$\frac{\partial R}{\partial U_{ay}}$

– Elements 1-10: partials w.r.t. U_{ay}

`dRdU`$_{ay}$`: G(234,ω,θ) – G(age`$_{ay}$`,ω,θ)`

$\frac{\partial R}{\partial \omega}$

– Element 11: partial w.r.t. ω

`dRdω:` `sum(U * (dGdω(234,ω,θ) – dGdω(age`$_{ay}$`,ω,θ)))` where

`dGdω(x,ω,θ):` `G(x,ω,θ) * (θ^ω) / (x^ω+θ^ω) * log(x/θ)`

$\frac{\partial R}{\partial \theta}$

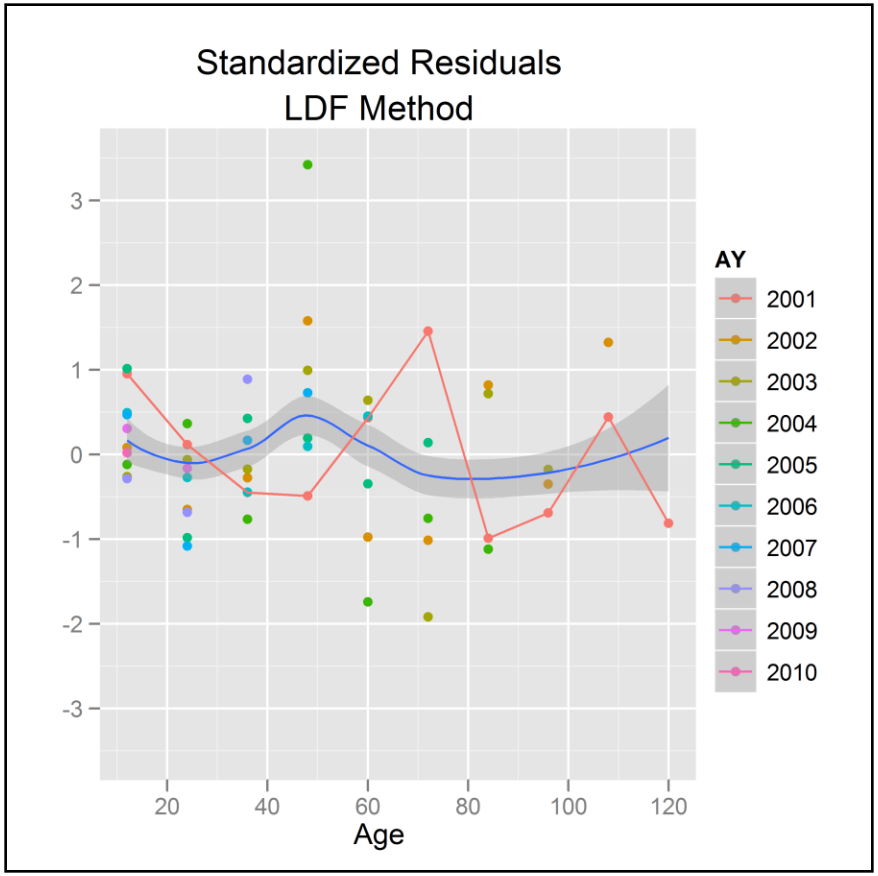
– Element 12: partial w.r.t. θ

`dRdθ:` `sum(U * (dGdθ(234,ω,θ) – dGdθ(age`$_{ay}$`,ω,θ)))` where

`dGdθ(x,ω,θ):` `G(x,ω,θ) * (θ^ω) / (x^ω+θ^ω) * (-ω/θ)`

- Hessian `H`: 12x12 matrix
  - R formulas for 2nd partial derivates are more complicated (follow Appendix A)
  - Or use Hessian from `optim`
    - Exercise care so numerical version of Hessian inverts
      - Problems when parameters (U, ω, θ) are scaled many orders of magnitude differently (slide 9)
    - Try dividing all loss values by 1 million before analysis begins

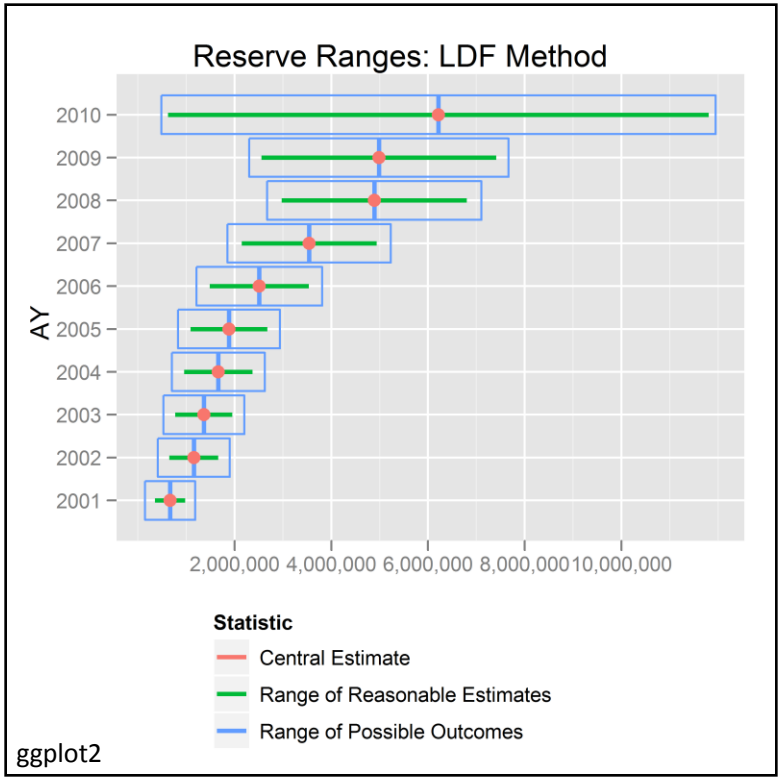
Wise to Inspect Residuals
May suggest inadequacies of, improvements to model



ggplot2

- Graph of average residual by age (blue curve, gray band) is not flat
  - Conclusion: loglogistic curve does not perfectly explain loss development by age
    - Not a surprise
  - Sinusoidal shape could be evidence of negative serial correlation in loss development
    - Suggests potential model "tweak"
- Only one outlier (|residual|>2)
  - Expect 2-3 = ~5% of 55 data points
  - Does not refute reasonableness of ODP assumption
- AY 2001 (jagged line) looks different here too
  - If decide to omit AY 2001 "subject" consider keeping mature observation(s)

Most Risk Due to Parameter Uncertainty
Reduce risk with more/different information



Range bands show ± 2 standard errors (SE's from slide 10)	
Red dot:	Central estimate
Green line:	Two-Std-Error Range of Reasonable Estimates
Blue rectangle:	Two-Std-Error Range of Possible Outcomes

- Larger sample size → smaller parameter risk
  - **Central Limit Theorem**: parameter estimates approach true values as sample size increases
  - 55 is a small number of data points, especially relative to 12 parameters
  - Does company data exist to analyze losses at more detailed (e.g., state, region, examiner, claim) levels?
- Incorporate supplemental data
  - Clark illustrates with hypothesized premium
    - A model that also reflects premium is *Cape Cod*
  - Point estimate for total does not change much
    - LDF Method $29.0 million
    - Cape Cod 29.7 million
  - *But total risk standard error drops significantly!*
    - LDF Method $ 4.9 million
    - Cape Cod 3.4 million
  - Benefit of more information from risk standpoint
    - ***Would it cost less than $1.5 million to collect premium?***
    - Enhance with sensitivity analysis

Why Consider Clark's Model?

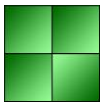
Actuarial models can be valuable to Management

- Models can produce point estimates with associated ranges
  - Requested by principals for years
- Models can improve risk analysis transparency
  - Models make explicit the connection between risk metrics and source data
  - Associated graphics help guide selection of which data to analyze
  - Modeling the right data improves
    - accuracy of estimates
    - relevancy of results
    - buy-in of message
- Models can add a risk dimension to cost/benefit analysis when requesting additional data

Statistical software environments – R or otherwise – are indispensible for the algorithms of today's sophisticated P&C models, such as those in Clark's innovative paper from 2003

- R is available for free at www.cran.r-project.org
- ggplot2, reshape, ChainLadder and many other packages available via download from that site
- Authors' request in return

Give credit where credit is due



Daniel Murphy, FCAS, MAAA
dmurphy@trinostics.com
www.trinostics.com