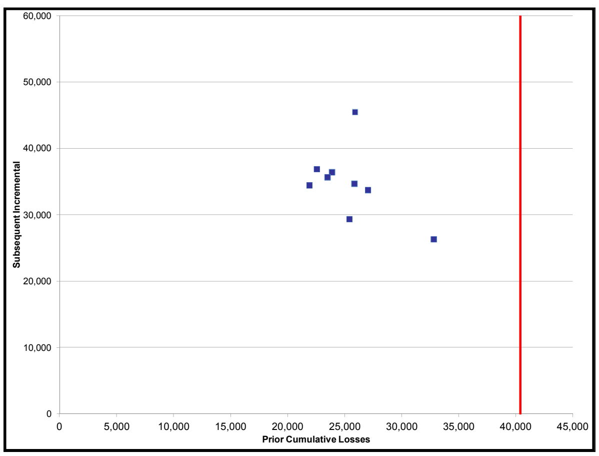
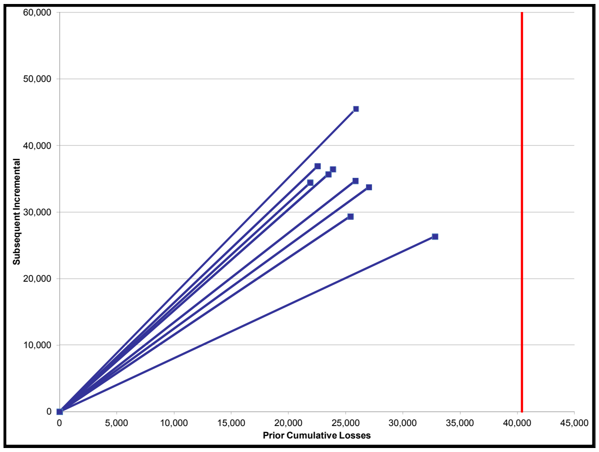
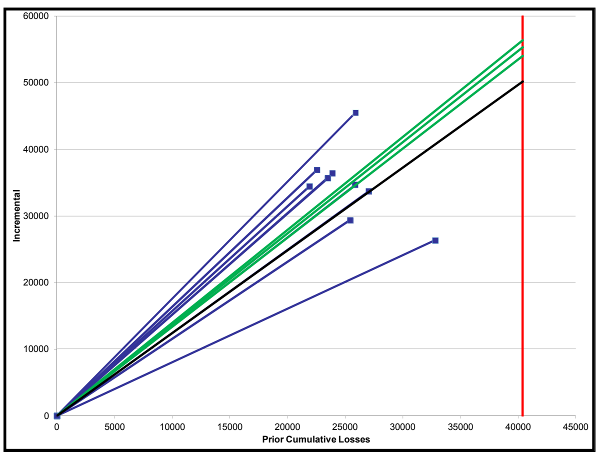
Heteroskedasticity in Loss Reserving

CASE Fall 2012

Cumulative Paid @12 Months	Cumulative Paid @24 Months	12-24 month LDF
21,898	56,339	2.5728
22,549	59,459	2.6369
23,881	60,315	2.5256
25,897	71,409	2.7574
23,486	59,165	2.5192
27,029	60,778	2.2486
25,845	60,543	2.3425
25,415	54,791	2.1559
32,804	59,141	1.8029





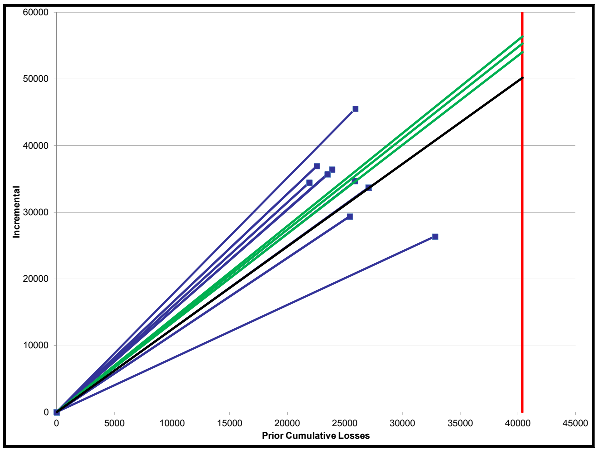


Method	LDF
Simple average	2.3958
Weighted average	2.3686
Unweighted least squares	2.3375

All of those estimators are unbiased.
Which one is efficient?

You're not weighting link ratios.

You're making an assumption about the variance of the observed data.



So how do we articulate our variance assumptions?

AY	Lag	CY	EarnedPremium	OnLevelEarnedPremium	CurrentPaid	PriorPaid	IncrementalPaid	CurrentIncurred	IncrementalIncurred	PriorIncurred
2001	1	2001	3166997	5555872	279455.0	0.0	279455.01	534986.6	534986.60	0.0
2001	2	2002	3166997	5555872	610822.7	279455.0	331367.74	2306840.7	1771854.12	534986.6
2001	3	2003	3166997	5555872	2288312.8	610822.7	1677490.09	2808189.6	501348.85	2306840.7
2001	4	2004	3166997	5555872	3088238.6	2288312.8	799925.73	3416244.9	608055.34	2808189.6
2001	5	2005	3166997	5555872	3341742.0	3088238.6	253503.46	3591742.0	175497.12	3416244.9
2001	6	2006	3166997	5555872	3341742.0	3341742.0	0.00	3441742.0	-150000.00	3591742.0
2001	7	2007	3166997	5555872	3341742.0	3341742.0	0.00	3441742.0	0.00	3441742.0
2001	8	2008	3166997	5555872	3341742.0	3341742.0	0.00	3341742.0	-100000.00	3441742.0
2001	9	2009	3166997	5555872	3341742.0	3341742.0	0.00	3341942.0	200.00	3341742.0
2001	10	2010	3166997	5555872	4680974.5	3341742.0	1339232.46	4680974.5	1339032.46	3341942.0
2001	11	2011	3166997	5555872	4690350.3	4680974.5	9375.80	4690350.3	9375.80	4680974.5
2002	1	2002	4512995	7150903	290663.2	0.0	290663.19	877172.1	877172.09	0.0
2002	2	2003	4512995	7150903	595625.8	290663.2	304962.60	1317474.3	440302.23	877172.1
2002	3	2004	4512995	7150903	1105174.2	595625.8	509548.43	1947943.8	630469.46	1317474.3
2002	4	2005	4512995	7150903	2491052.6	1105174.2	1385878.36	2692664.2	744720.40	1947943.8
2002	5	2006	4512995	7150903	2622311.8	2491052.6	131259.19	2919485.3	226821.14	2692664.2
2002	6	2007	4512995	7150903	2780267.5	2622311.8	157955.76	3040422.6	120937.24	2919485.3
2002	7	2008	4512995	7150903	2963319.9	2780267.5	183052.41	3095819.9	55397.38	3040422.6

- A variable of interest (paid losses, for example) presumed to have some statistical relationship to one or more other variables in the matrix.
- The strength of that relationship may be established by creating models which relate two variables.
- A third variable is introduced by categorizing the predictors.
- Development lag is generally used as the category.

The response variable will generally be incremental paid or incurred losses.

We assume that error terms are homoskedastic and normally distributed

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix}$$

The design matrix may be either prior period cumulative losses, earned premium or some other variable. Columns are differentiated by category.

Calibrated model factors are analogous to age-to-age development factors.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix} =$$

21,898	56,339	78,457	92,820	101,261	103,929	107,855	111,901	112,388	111,727
22,549	59,459	82,918	98,768	105,504	110,649	113,833	116,691	116,055	
23,881	60,315	85,828	98,656	106,590	110,107	113,159	113,243		
25,897	71,409	95,443	111,051	120,780	124,499	124,352			
23,486	59,165	81,511	96,077	107,063	109,505				
27,029	60,778	80,161	89,707	92,374					
25,845	60,543	76,743	82,470						
25,415	54,791	64,854							
32,804	59,141								
40,409									

Loss Reserving &
Ordinary Least Squares Regression

A Love Story

LSM	$y = bx + e$
WAD	$y = bx + \sqrt{x}e$
SAD	$y = bx + xe$

Murphy: Unbiased Loss Development Factors

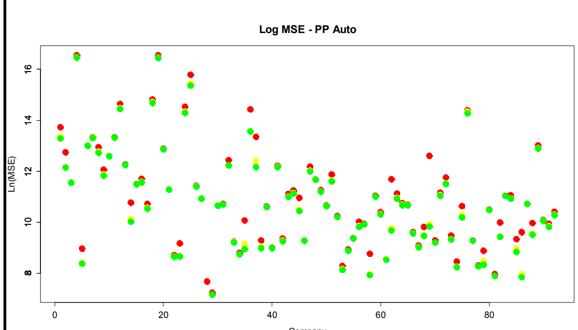
$$y = bx + x^{\alpha/2}e$$

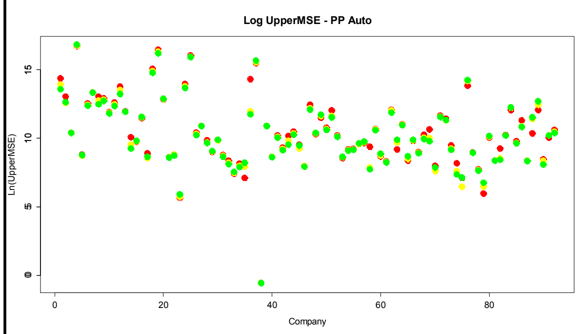
The multivariate model may be generally stated as containing a parameter to control the variance of the error term.

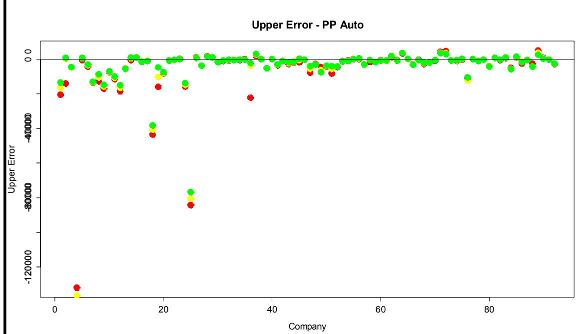
α is not a hyperparameter. Fitting using SSE will always return α = 0

- Intuition
  - Losses vary in relation to predictors
  - Loss ratio variance looks different
- Observation
  - Behavior of a population
  - Diagnostics on individual sample (Breusch-Pagan test)

In 2011, Glenn Meyers & Peng Shi published NAIC Schedule P results for 132 companies. The object was to create a laboratory to determine which loss reserving method was most reliable.







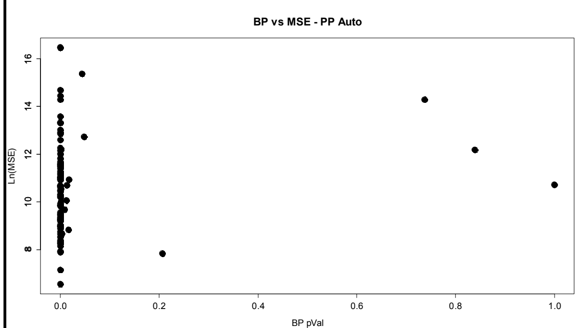
THE REDWOODS GROUP

Breusch-Pagan Test

- Use regression to diagnose your regression
- Does the variance depend on the predictor?
- Regress squared residuals against the predictor

$$e_i^2 = \beta_0 + \beta_1 x_i$$

An F-test determines the probability that the coefficients are non-zero.



Caveats

- Non-normal error terms render B-P meaningless!
- Chain ladder utilizes stochastic predictors
- Earned premium has not been adjusted

- Breusch-Pagan test is not strongly persuasive across the total data set.
- Homoskedastic error terms would support unweighted calibration of model factors.
- Probably more important to test functional form of error terms. Kolmogorov-Smirnov etc. may test for normal residuals.

State your model and your underlying assumptions.

Test those assumptions.

Stop using models whose assumptions don't reflect reality!

Statisticians have been doing this for years. Easy to ~~steal~~ leverage their work.

“Abandon your triangles!”

-Dave Clark
CAS Forum 2003

THE REDWOODS GROUP

- https://github.com/PirateGrunt/CASE-Spring-2013
- PirateGrunt.com
- http://lamages.blogspot.com/