CREDIBILITY-LIKE SHRINKAGE IN LINEAR MODELS FOR PRICING AND RESERVING

Gary G Venter

CAS Annual Meeting

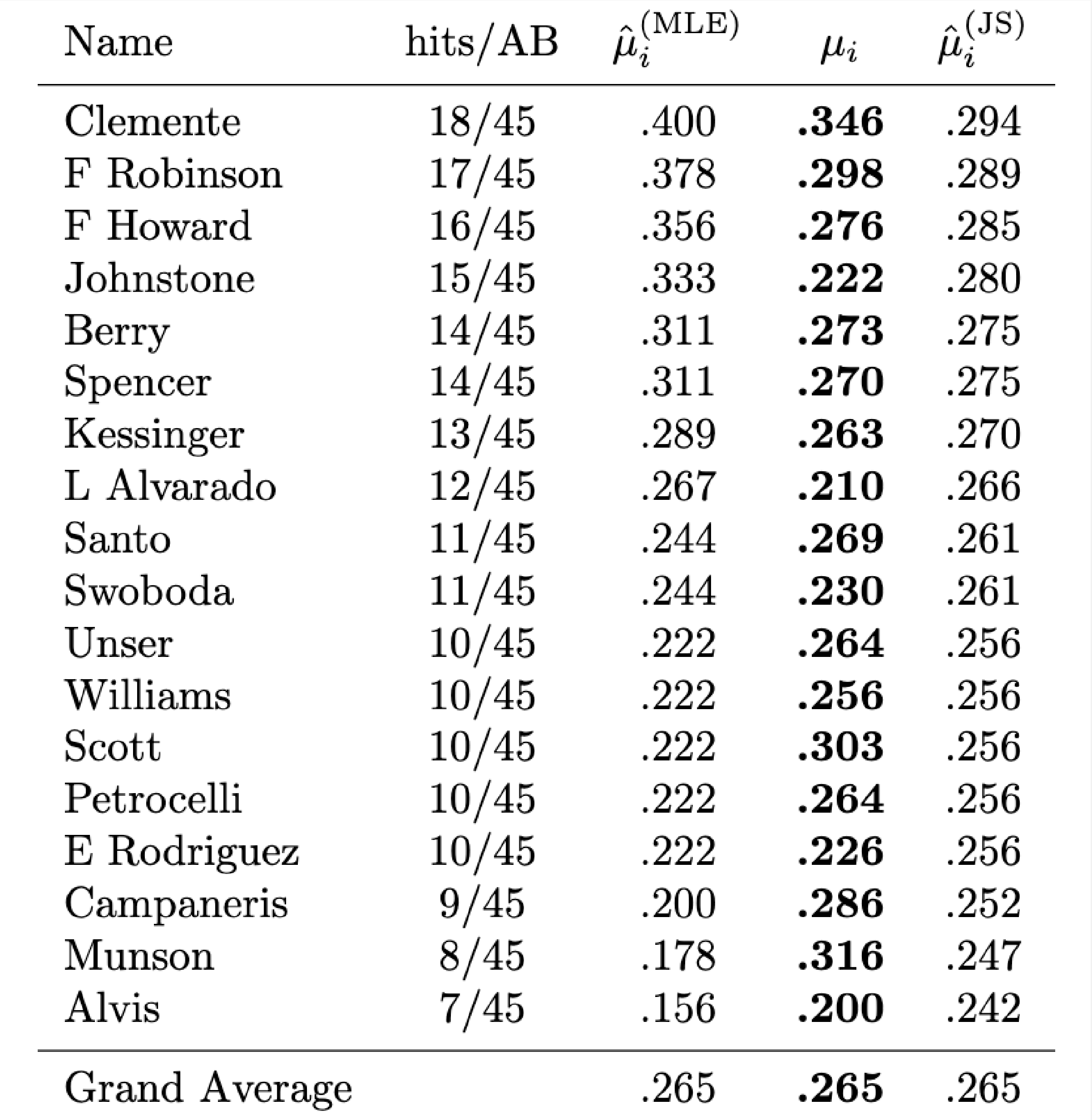
Honolulu November 2019

CREDIBILITY REDUCES PREDICTIVE VARIANCE

- Actuarial credibility minimizes sum of squared errors between estimates and true mean
  - Also minimizes variance of **prediction** errors; Stein’s Theorem from 1959 says some degree of weighting towards mean always reduces error variance when 3 or more cells being estimated
  - Shrinks estimates towards overall mean, so estimates are biased towards grand mean, but estimation errors are reduced –error bands are smaller but no longer symmetric
- Buhlmann’s least squares credibility from 1968 same as the James-Stein estimator from 1961
  - They assume normal distributions and he minimizes squared errors and calls it non-parametric
  - But least squares = MLE for normal distributions, so both are really making the same assumptions. Also credibility doesn’t work well unless distributions close to normal.
- Credibility factor Z from ratio of variance components. In actuarial terminology this is: expected process variance / variance of hypothetical means
- But James-Stein considers cases where you don’t know the variance of the hypothetical means (though you could get close from historical baseball statistics in the example)
  - Estimate that as sample variance*(N – 1)/(N – 3) when there are N means being estimated

BATTING AVERAGE EXAMPLE

- Used in many papers by Efron, Morris, Van Slyke… This one from http://statweb.stanford.edu/~ckirby/brad/LSI/chapter1.pdf
- 18 MLB players after 45 at bats in 1970
- Estimate true average for each – to be measured by end of season average 350 or so at bats later
- Credibility weights each average against overall mean with Z = 0.212
- A lot of shrinkage towards mean, as early average volatile
- Reduces predictive variance by a factor of 3.5, so to 28% of what it would be by MLE – which is predicting each mean by its early-season average



Name	hits/AB	$\hat{\mu}_i^{(\text{MLE})}$	μ_i	$\hat{\mu}_i^{(\text{JS})}$
Clemente	18/45	.400	**.346**	.294
F Robinson	17/45	.378	**.298**	.289
F Howard	16/45	.356	**.276**	.285
Johnstone	15/45	.333	**.222**	.280
Berry	14/45	.311	**.273**	.275
Spencer	14/45	.311	**.270**	.275
Kessinger	13/45	.289	**.263**	.270
L Alvarado	12/45	.267	**.210**	.266
Santo	11/45	.244	**.269**	.261
Swoboda	11/45	.244	**.230**	.261
Unser	10/45	.222	**.264**	.256
Williams	10/45	.222	**.256**	.256
Scott	10/45	.222	**.303**	.256
Petrocelli	10/45	.222	**.264**	.256
E Rodriguez	10/45	.222	**.226**	.256
Campaneris	9/45	.200	**.286**	.252
Munson	8/45	.178	**.316**	.247
Alvis	7/45	.156	**.200**	.242
Grand Average		.265	**.265**	.265

EXTENDING THIS TO REGRESSION

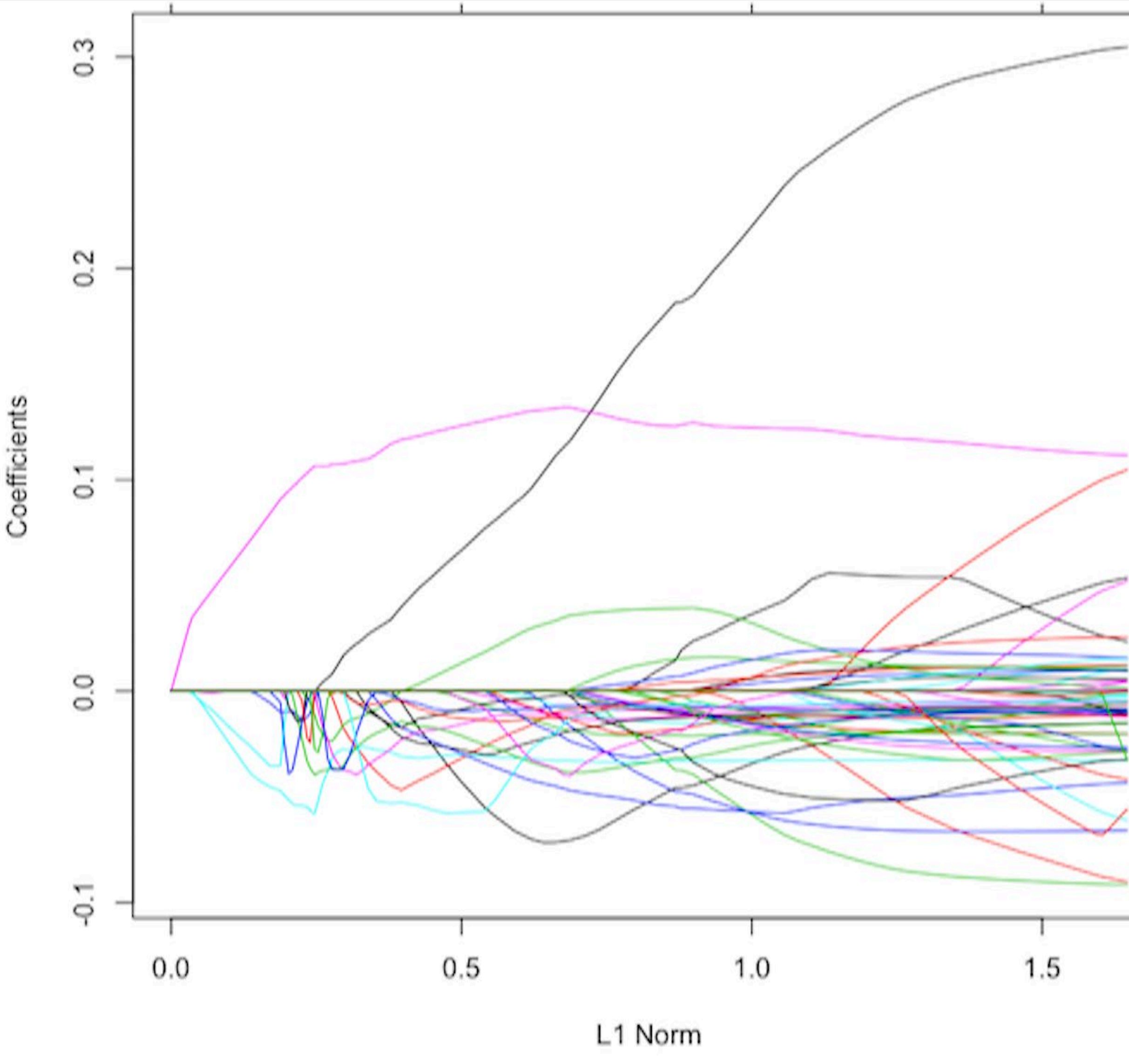
- One direction for credibility for regression is Charles Hachemeister's paper on that from the 1974 Berkeley credibility conference, reproduced at https://www.casact.org/pubs/forum/92spforum/92sp307.pdf
- I worked for Charlie starting late 1977 and he trained me to be a research actuary. I was at that conference as well as a new actuarial trainee but didn't understand any of it.
- My 2008 ASTIN Bulletin paper with Jose Couret on workers comp hazard group frequency by injury type was an application of this. See https://www.casact.org/library/astin/vol38no1/73.pdf
- We had data on injury-type frequency by hazard group over time, so within and between variances and covariances in all directions, and could apply credibility to weight injury-type frequencies with those in the other hazard groups and to other injury types in that group.
- However in usual regressions you do not have all those variances. Other approaches have been developed for shrinkage of fitted values towards the mean to reduce error variances, based on shrinking the regression coefficients. Theorem: ∃ some degree of shrinkage that is better than none

SHRINKAGE REDUCES PREDICTIVE VARIANCE IN LINEAR MODELS

- 1970 paper by Hoerl and Kennard introduced ridge regression
- This minimizes negative loglikelihood (NLL) plus λ*sum of squared parameters, for some factor λ
- Pushes parameters closer to zero, depending on how much each one improves the NLL
- But first you standardize all variables by subtracting their means from each, and dividing all regressor variables by their standard deviations to make their scales comparable
- Add back mean of dependent variable to the regression estimate of its differences from mean
- All fitted differences from mean are shrunk towards zero as they are linear combinations of mean zero variables, and their coefficients have been shrunk. So biased but with less error.
- All fitted values are shrunk toward the overall mean, just like in credibility – and best λ always > 0
- Now it is common not to standardize the dependent variable, and then not to shrink the constant
- Select λ by cross-validation: leave out maybe a rotating 10% of the data in 10 separate regressions, and measure NLL of the left out parts, then add these up to give a penalized NLL – look for λ that minimizes that penalized NLL

LASSO (LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR)

- Minimize NLL + $\lambda\Sigma_j|\beta_j|$ for parameters β_j.
- First appeared in *Santosa, Fadil; Symes, William W. (1986)*, but reinvented and popularized by *Tibshirani, Robert (1996)*.
- Using absolute values shrinks some parameters to exactly zero
- That's why the term selection
- Can start off with a lot of parameters and it gets the best combination of them for each λ.
- When fitting you get a graph like this that shows how the coefficients of the model increase as L1 norm (i.e., $\Sigma|\beta_j|$) increases and λ decreases.
- In this case, parameters are negatively correlated so they come in and out of the model



RANDOM EFFECTS

- Assume parameters (i.e., “effects”) are (typically) iid normal $(0, \sigma^2)$
- Maximize joint likelihood, which is probability of parameters times the likelihood, where likelihood = the probability of data given parameters. Joint likelihood is then the joint density of the parameters and the data, by definition of conditional distribution.
- From normal distribution, – log of probability of parameters β_j, for fixed σ^2, is: (some constant) + $\Sigma_j \beta_j^2/2\sigma^2$.
- Maximizing the joint likelihood means minimizing NLL + $\Sigma_j \beta_j^2/2\sigma^2$, so is ridge regression with $\lambda = 1/2\sigma^2$.
- But if σ^2 is not fixed, it gets estimated as well when maximizing joint likelihood. A way to estimate λ.
- If parameters are double exponential (i.e., $|\beta_j|$ are exponential), this instead produces lasso. Called Bayesian lasso for reasons coming.

BAYESIAN VERSION

- Joint likelihood is also the probability of the data times the probability of the parameters given the data, again by the definition of conditional distribution
- Probability of the data is an unknown constant. Thus joint likelihood is proportional to the conditional distribution of the parameters given the data
- MCMC (Markov Chain Monte Carlo) is a way to generate a sample of a distribution if it is known up to a constant. It has to integrate to 1.0.
- So it can approximate the probability distribution of the parameters given the data – which is what Bayesians call the posterior. This terminology is a bit antiquated, as doing it does not require subjective probability or Bayes Theorem, and we can do it all as frequentist random effects
- The ridge regression or lasso estimate from maximizing the joint likelihood is the mode of the conditional distribution of the parameters given the data. Frequentists routinely calculate this mode so why not the whole distribution?

PENALIZED LIKELIHOOD MEASURING GOODNESS OF FIT

- Good penalized likelihood measures are the small sample AIC – denoted AICc – and the HQIC – Hannan-Quinn Information Criterion
- They add a penalty to the NLL. With sample size N and k parameters, penalties are:
  - AICc: $k * N/[N - k - 1]$ HQIC: $k * \ln[\ln(N)]$
- Goal is to take out the sample bias in the NLL, so the penalized NLL would be the NLL for a new, independent sample when using the parameters fit to this sample
- But shrunk parameters don't act like full parameters – they use up fewer degrees of freedom – so we don't know the right k to use
- Can use cross validation to do penalized likelihood – leave out say 1/5 of sample for fitting, get NLL of the left out points and repeat 4 times and add up left out NLLs
- Or subtract the penalty from LL
- Extreme case of this is leave-one-out (loo), where each point is used by itself as the left-out subsample

LOO

- Sounds like a lot of work
- But if you have the whole conditional distribution of the parameters given the data, there is a good, simple approximation
- Estimate the left-out likelihood of a point as a weighted average of its likelihoods across the sample of parameter sets, where worse-fitting samples get more weight
- A technique called importance sampling gives each parameter set a weight inversely proportional to the point's likelihood using those parameters. The left-out estimate is then the harmonic mean of the likelihoods of the point.
- This turns out to be a volatile estimate. Using a kind of extreme-value adjustment for the very bad reciprocal likelihoods gets an improvement called Pareto-smoothed importance sampling. There is software in R to do this.
- This is a good estimate, and gives a good estimate of LL adjusted for sample bias.

USING LOO

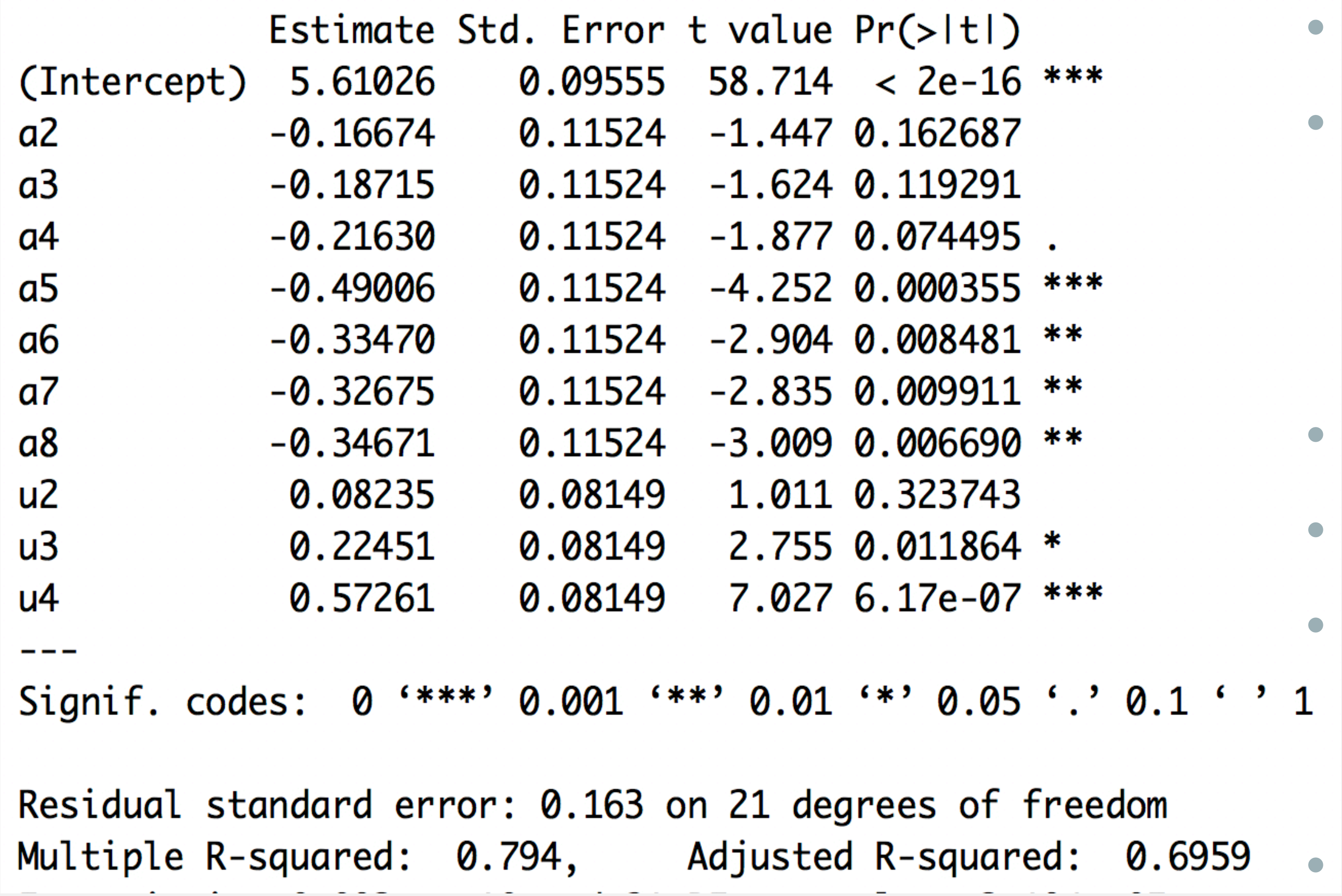
- The penalized LL – just called loo – gives fit comparisons for different models.
- Can be used to optimize λ.
- But a problem is there is some random estimation error in the bias adjustment
- Maximizing loo by a search over all parameters will very likely result in a model where loo is over-estimated. Different than picking best loo from a few choices.
- An alternative is to make the random-effects σ also a random effect – or in Bayesian terms, specify a prior for σ. Sometimes called a hyper prior.
- Choices of this prior seem not to make much of a difference in the final model.
- Usually the resulting σ – and so λ – gets as good a loo as any λ does – and often a slightly better one, apparently resulting from having a distribution of λs
- Called the fully-Bayesian estimate – optimizing loo isn't a Bayesian step
- Also is fully frequentist as all the so-called parameters are random effects – there are no fixed effects in this approach – i.e., no parameters in the frequentist sense

COLLISION SEVERITY BY AGE AND USE FU / WU VARIANCE PAPER 2007

- 8 age of driver classes, 4 use classes: pleasure, drive to work short, drive to work long, business use
- Log of severity regressed for by age and use class
- Exponentiating the model gives a multiplicative model
- Regression uses 0,1 dummy variables for all but 1st age and 1st use: age 1, use 1 is the constant term
- Other cells: log severity = constant + age effect + use effect + ε. Then severity = Age effect * Use effect*e^{ε}
- Fitted value is sumproduct of parameter vector and vector of dummies for the point plus the constant term
- In matrix notation, let y be the column vector of log severities shown, x be the design matrix, b be the vector of parameters, and c the constant
- Then fitted value vector $\underline{y}$ = c + xb

Age	Use	ln_s	a2	a3	a4	a5	a6	a7	a8	u2	u3	u4
1	1	5.52	0	0	0	0	0	0	0	0	0	0
1	2	5.62	0	0	0	0	0	0	0	1	0	0
1	3	5.50	0	0	0	0	0	0	0	0	1	0
1	4	6.68	0	0	0	0	0	0	0	0	0	1
2	1	5.36	1	0	0	0	0	0	0	0	0	0
2	2	5.70	1	0	0	0	0	0	0	1	0	0
2	3	5.70	1	0	0	0	0	0	0	0	1	0
2	4	5.89	1	0	0	0	0	0	0	0	0	1
3	1	5.52	0	1	0	0	0	0	0	0	0	0
3	2	5.52	0	1	0	0	0	0	0	1	0	0
3	3	5.70	0	1	0	0	0	0	0	0	1	0
3	4	5.84	0	1	0	0	0	0	0	0	0	1
4	1	5.43	0	0	1	0	0	0	0	0	0	0
4	2	5.43	0	0	1	0	0	0	0	1	0	0
4	3	5.68	0	0	1	0	0	0	0	0	1	0
4	4	5.91	0	0	1	0	0	0	0	0	0	1
5	1	5.03	0	0	0	1	0	0	0	0	0	0
5	2	5.31	0	0	0	1	0	0	0	1	0	0
5	3	5.47	0	0	0	1	0	0	0	0	1	0
5	4	5.55	0	0	0	1	0	0	0	0	0	1
6	1	5.34	0	0	0	0	1	0	0	0	0	0
6	2	5.31	0	0	0	0	1	0	0	1	0	0
6	3	5.46	0	0	0	0	1	0	0	0	1	0
6	4	5.87	0	0	0	0	1	0	0	0	0	1
7	1	5.34	0	0	0	0	0	1	0	0	0	0
7	2	5.31	0	0	0	0	0	1	0	1	0	0
7	3	5.54	0	0	0	0	0	1	0	0	1	0
7	4	5.83	0	0	0	0	0	1	0	0	0	1
8	1	5.26	0	0	0	0	0	0	1	0	0	0
8	2	5.28	0	0	0	0	0	0	1	1	0	0
8	3	5.56	0	0	0	0	0	0	1	0	1	0
8	4	5.84	0	0	0	0	0	0	1	0	0	1

FIRST A STRAIGHT REGRESSION ON LOGS WITH T-STATISTICS



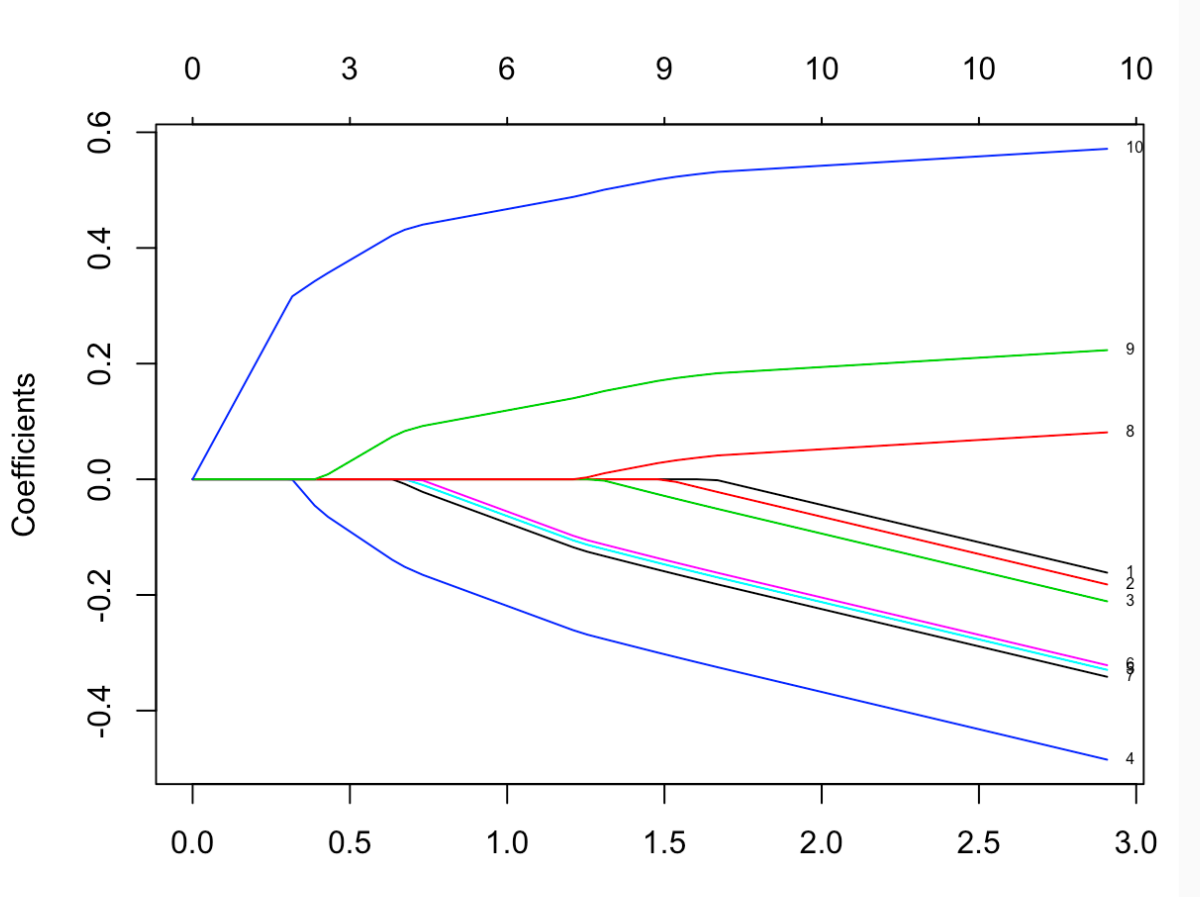
```
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.61026    0.09555  58.714  < 2e-16 ***
a2          -0.16674    0.11524  -1.447 0.162687    
a3          -0.18715    0.11524  -1.624 0.119291    
a4          -0.21630    0.11524  -1.877 0.074495 .  
a5          -0.49006    0.11524  -4.252 0.000355 ***
a6          -0.33470    0.11524  -2.904 0.008481 ** 
a7          -0.32675    0.11524  -2.835 0.009911 ** 
a8          -0.34671    0.11524  -3.009 0.006690 ** 
u2           0.08235    0.08149   1.011 0.323743    
u3           0.22451    0.08149   2.755 0.011864 *  
u4           0.57261    0.08149   7.027 6.17e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.163 on 21 degrees of freedom
Multiple R-squared:  0.794,	Adjusted R-squared:  0.6959 
```

- Age 5 and use 4 most significant
- Ages below group 5 not at all so as well as use 2. Can leave out those variables, which mean they get the constant
- Older ages somewhat significant
- Use 3 a little less important
- R-squared is fraction of variance explained by the model. Adjusted is for number of parameters
- Fit is from R lm function

NOW TRY LASSO AND BAYESIAN LASSO

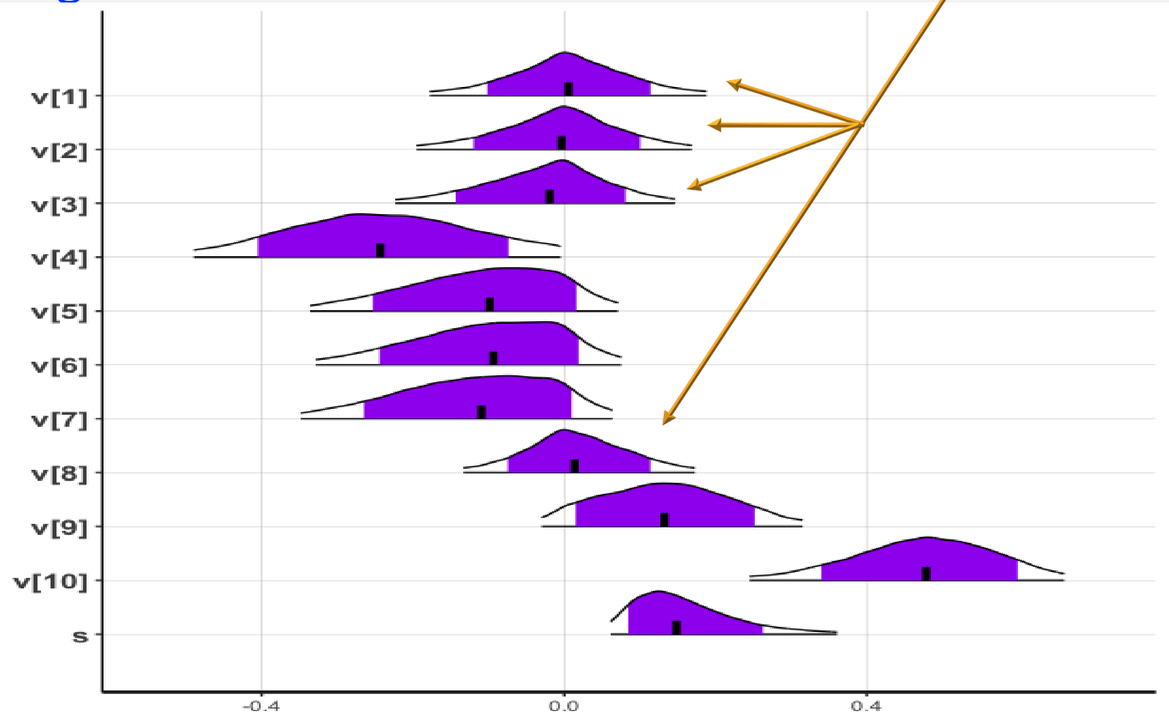
- Use R glmnet function for lasso
- Use rstan for MCMC for Bayesian lasso
  - Run in R but write regression function in rstan
- glmnet has a cross-validation function called cv.glmnet
- Produces estimates for a minimum λ called lambda.min and a larger one called lambda.1se. Can make one called Mid at their geometric mean.
- glmnet gives graph for parameter values as 1/λ increases and adds in parameters
- Parameters from these methods compared on next slide



PARAMETER COMPARISON

	Regr	Min	Mid	1 se	MCMC	t
c	5.61	5.49	5.47	5.48	5.49	58.7
a2	-0.17	.	.	.	0.01	-1.4
a3	-0.19	.	.	.	-0.01	-1.6
a4	-0.22	-0.02	.	.	-0.03	-1.9
a5	-0.49	-0.29	-0.21	-0.08	-0.24	-4.3
a6	-0.33	-0.14	-0.05	.	-0.11	-2.9
a7	-0.33	-0.13	-0.05	.	-0.10	-2.8
a8	-0.35	-0.15	-0.07	.	-0.12	-3.0
u2	0.08	0.02	.	.	0.02	1.0
u3	0.22	0.16	0.11	0.02	0.13	2.8
u4	0.57	0.51	0.46	0.37	0.47	7.0

- MCMC generally between Min and Mid here
- Used prior of uniform[-5,5] for log of double-exponential s, which is related to λ. Mean was -1.9. Uniform[-10,10] for log constant. Mean was 1.7.
- Can take out variables if mean near 0, spread big
- Stan gives graph of posteriors, here with 80% ranges

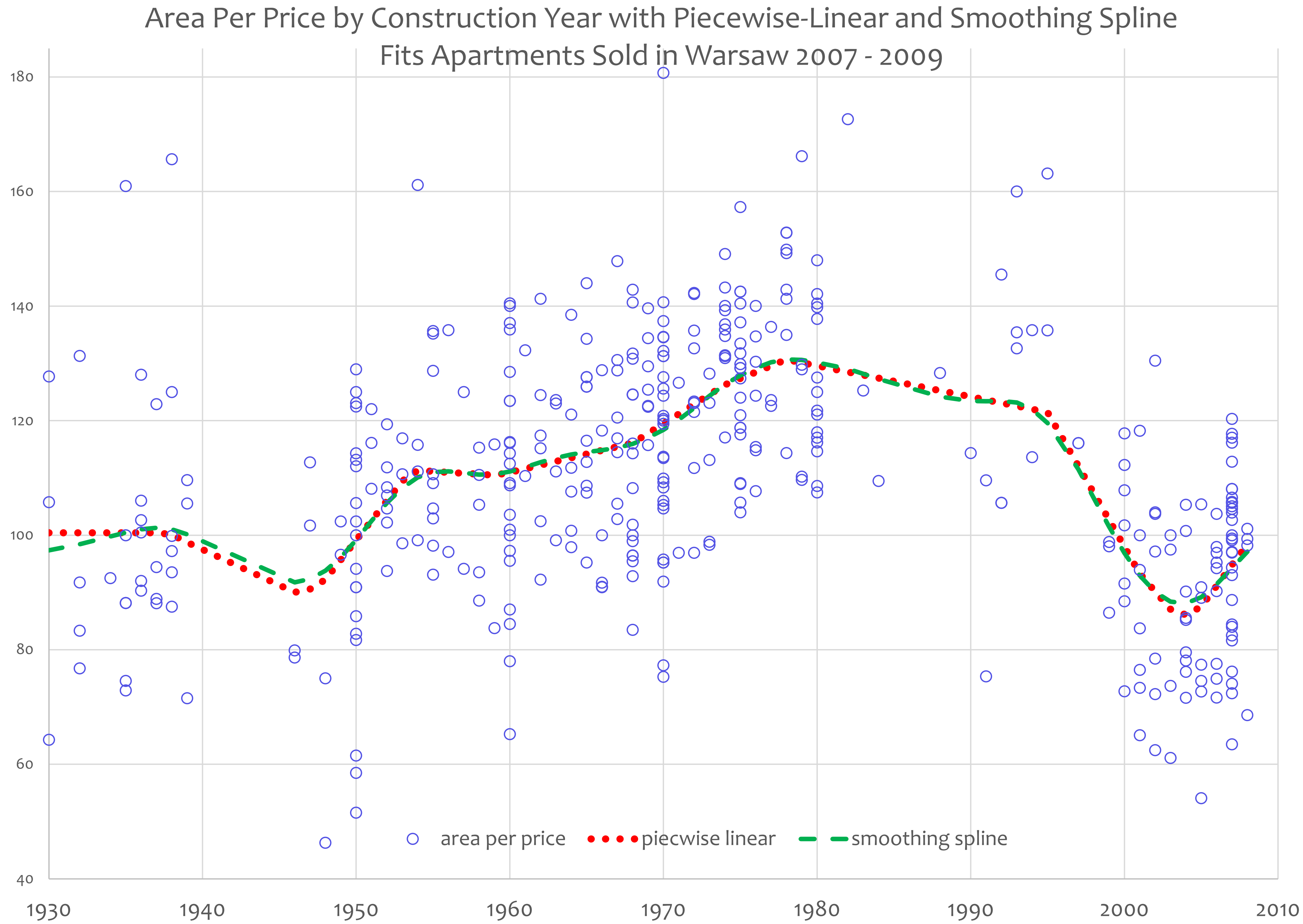


FITTING CURVES ACROSS PARAMETERS

- Actuaries have used for fitting curves to loss development parameters
- Now a big focus in statistics is building up curves in small sections
- Using line segments for this is called piecewise linear, or linear splines
- Cubic splines are used in some actuarial estimation as well
- O’Sullivan penalized cubic splines are generalizations that apply shrinkage to cubic splines – not closed form but functions to do it available in R
- These are the prime tool used in semi-parametric regression for putting these not-quite-parametric curves across the parameters – see: http://semiparametric-regression-with-r.net/
- Cross-validation used for determining how much shrinkage to use – also for selecting where to put knots that link the splines.
- Several steps and a lot of work – programs available now but kind of black box

ACTUARIAL SEMI-PARAMETRIC REGRESSION: SHRINKING SLOPE CHANGES OF LINEAR SPLINES

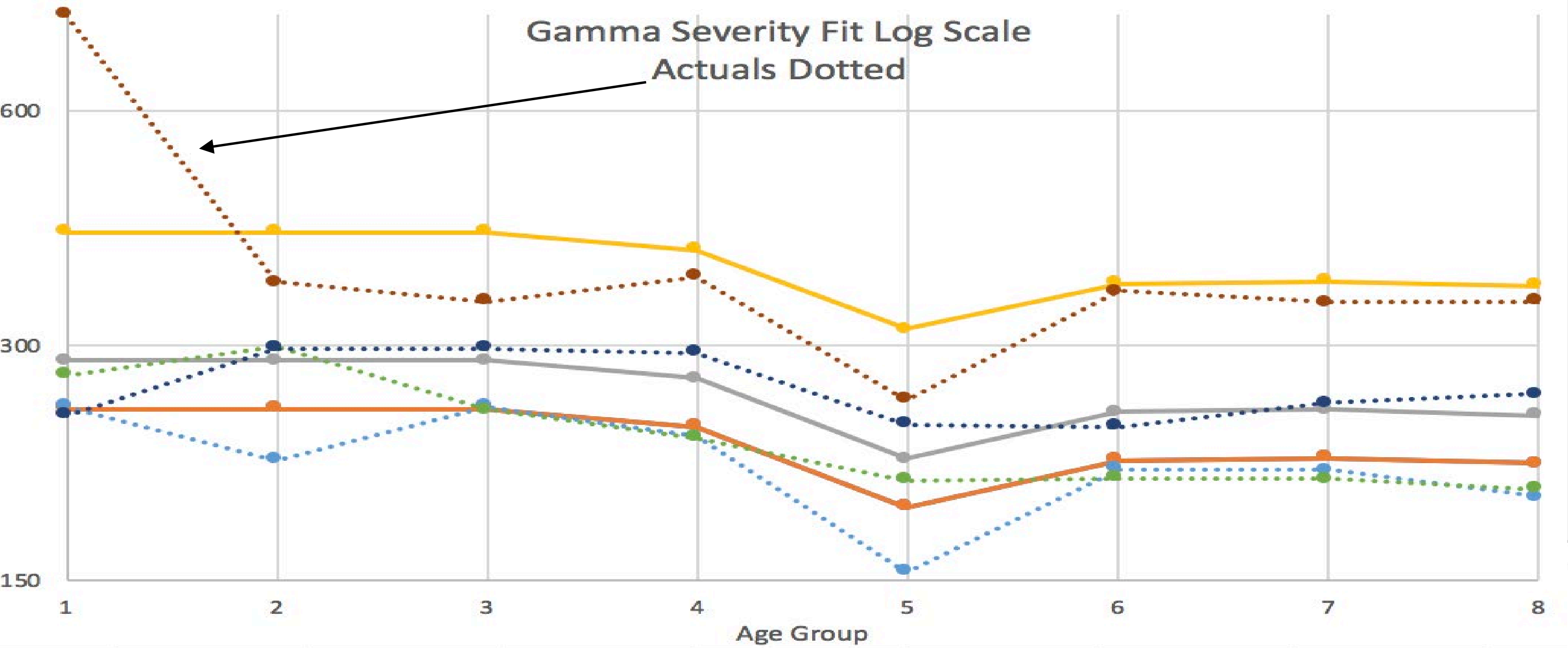
- Papers by Barnett, Zehnwirth, Gluck, Venter, Șahin, Gutkovich, Gao, …
- Originally used other forms of shrinkage but now use Bayesian shrinkage – including generalizations of Bayesian lasso with other shrinkage priors
- Set up as a regression with slope-change dummy variables, so parameter shrinkage shrinks the slope changes. If any of those goes to 0 then old slope continues. Thus also finds where the knots are needed.
- Cross-validation gives estimate of how well a model would fit a new sample from the same population but there is estimation error
- Automated optimization of a cross-validation measure has high risk of finding a model where the implied population fit is exaggerated
- Fully Bayesian approach avoids this. Simple and automatic procedure.

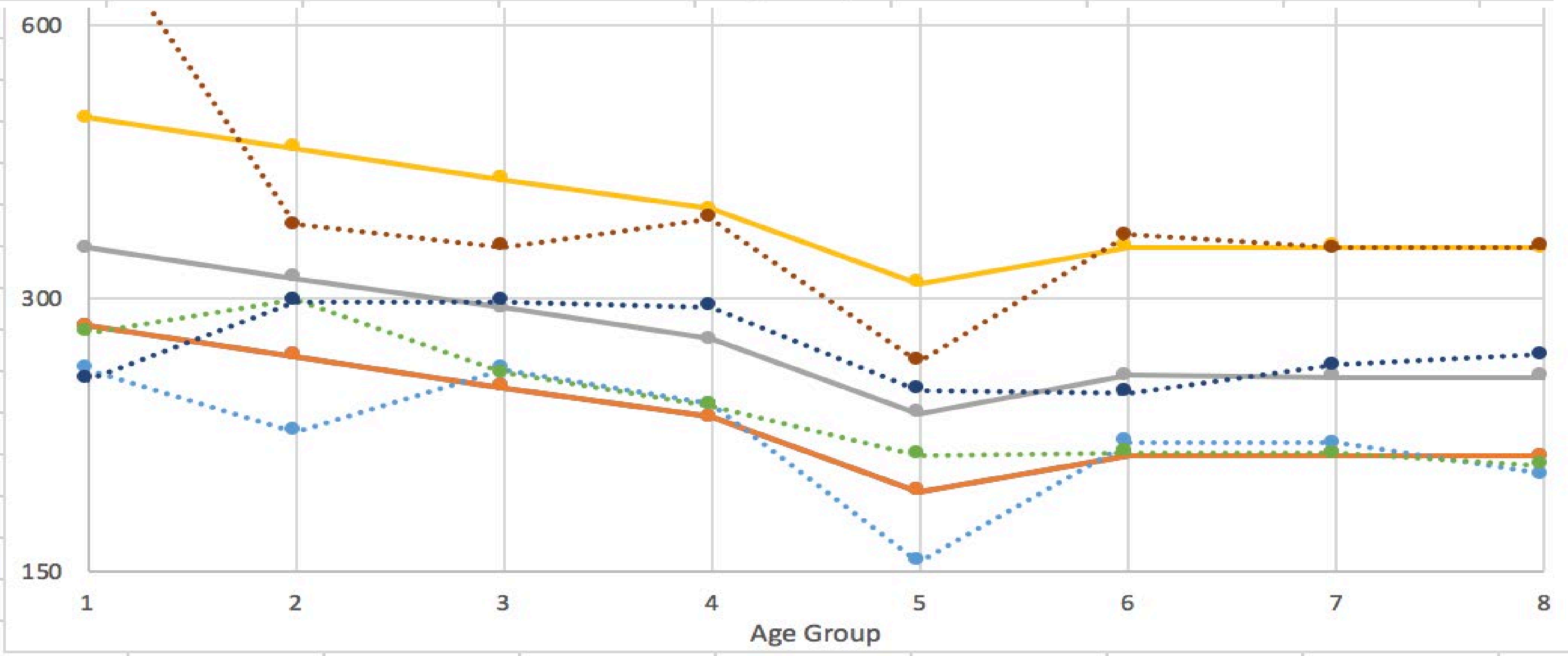


Example: Comparison of piecewise linear done in Stan vs. O’Sullivan splines from link above

Very similar fits

Can also be done to make curves across parameters in regressions





Fu/Wu collision severity data can use piecewise linear fits to any or all of the parameter types

Here it is done for age groups. Lines are for each use

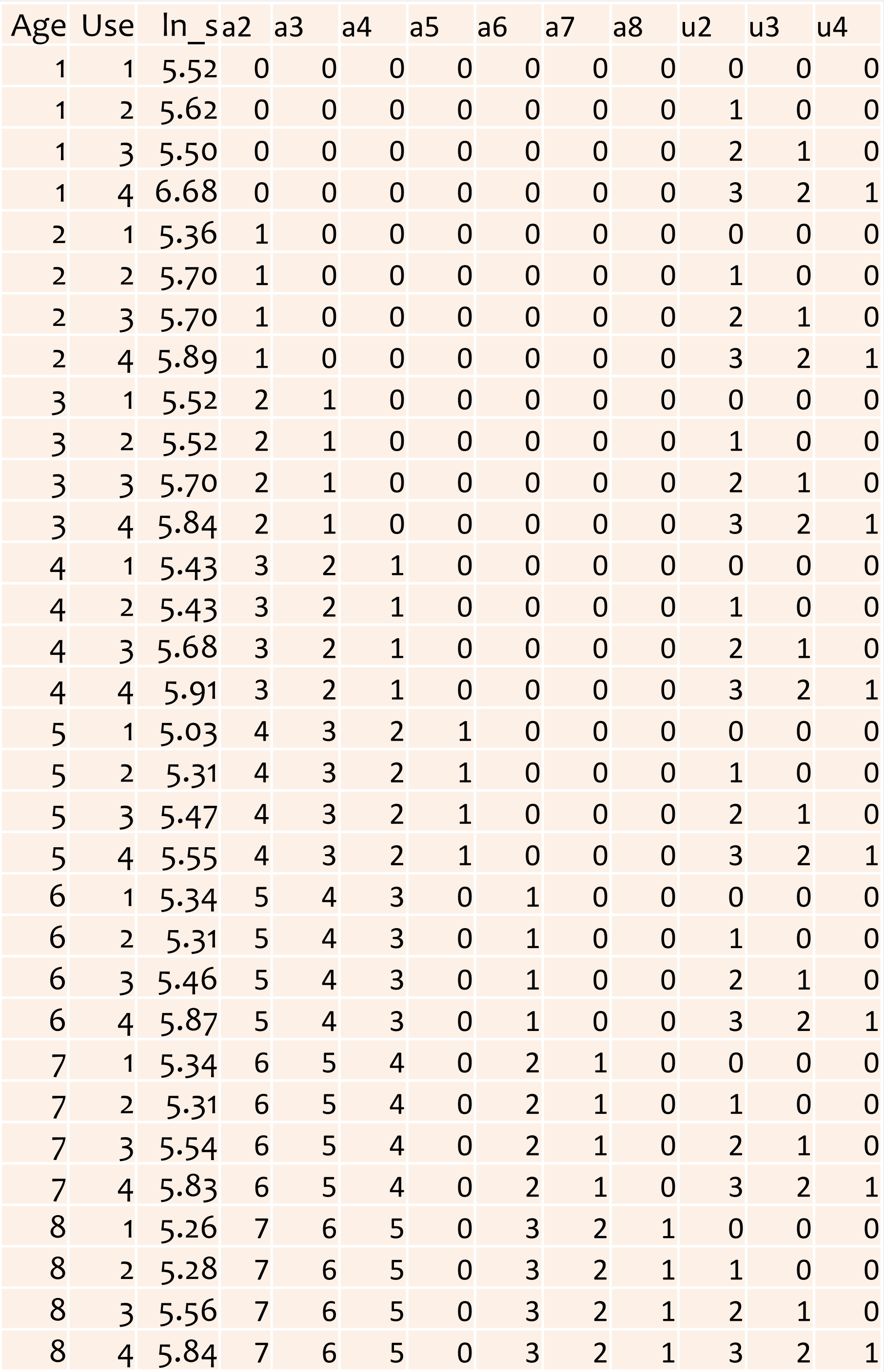
Top graph is original fit, with parameters by age and use. Bottom is with piecewise linear fit by ages. Note that the slope flattens out for the last 3 age groups

For both, pleasure use and short drive to work classes got same fit, shown in red, between the two actuals, which are the dotted lines

Using this kind of trend gave a bit better fit by loo.

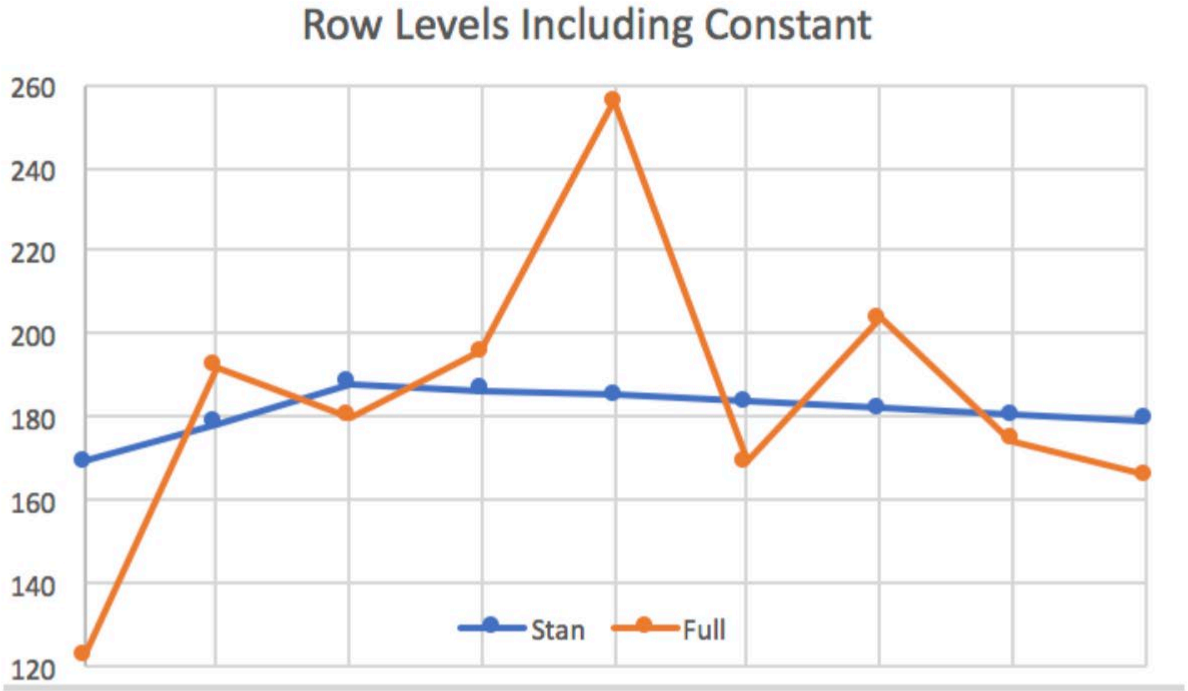
SLOPE CHANGE DUMMY VARIABLES

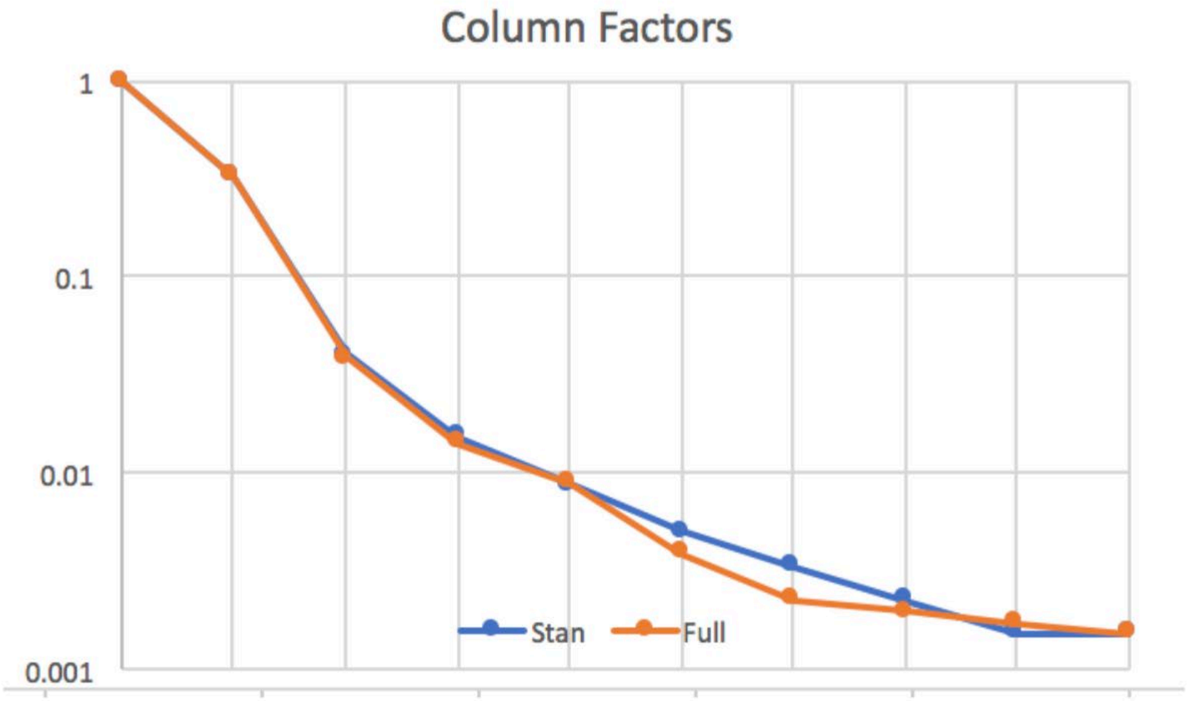
- Focus on dummy variable for age 2 slope change
- An observation from age 1 doesn't use this variable, so its dummy is 0 there
- The slope change for age 2 is the slope for age 1 to age 2, and is also the parameter for age 2, so the dummy is 1 at age 2.
- At age 3, the slope is the slope change for age 3 plus the slope at age 2. So the parameter is the parameter at age 2 plus this slope, so the contribution of the age 2 variable is 2.
- At age 4, the dummy is 3, etc.
- The dummy for age j at age i is max(0, 1 + i – j)
- Same for the use class slope change dummies.
- Here age 5 is not a slope change but is a separate parameter – modeling judgment call



Age	Use	ln_s	a2	a3	a4	a5	a6	a7	a8	u2	u3	u4
1	1	5.52	0	0	0	0	0	0	0	0	0	0
1	2	5.62	0	0	0	0	0	0	0	1	0	0
1	3	5.50	0	0	0	0	0	0	0	2	1	0
1	4	6.68	0	0	0	0	0	0	0	3	2	1
2	1	5.36	1	0	0	0	0	0	0	0	0	0
2	2	5.70	1	0	0	0	0	0	0	1	0	0
2	3	5.70	1	0	0	0	0	0	0	2	1	0
2	4	5.89	1	0	0	0	0	0	0	3	2	1
3	1	5.52	2	1	0	0	0	0	0	0	0	0
3	2	5.52	2	1	0	0	0	0	0	1	0	0
3	3	5.70	2	1	0	0	0	0	0	2	1	0
3	4	5.84	2	1	0	0	0	0	0	3	2	1
4	1	5.43	3	2	1	0	0	0	0	0	0	0
4	2	5.43	3	2	1	0	0	0	0	1	0	0
4	3	5.68	3	2	1	0	0	0	0	2	1	0
4	4	5.91	3	2	1	0	0	0	0	3	2	1
5	1	5.03	4	3	2	1	0	0	0	0	0	0
5	2	5.31	4	3	2	1	0	0	0	1	0	0
5	3	5.47	4	3	2	1	0	0	0	2	1	0
5	4	5.55	4	3	2	1	0	0	0	3	2	1
6	1	5.34	5	4	3	0	1	0	0	0	0	0
6	2	5.31	5	4	3	0	1	0	0	1	0	0
6	3	5.46	5	4	3	0	1	0	0	2	1	0
6	4	5.87	5	4	3	0	1	0	0	3	2	1
7	1	5.34	6	5	4	0	2	1	0	0	0	0
7	2	5.31	6	5	4	0	2	1	0	1	0	0
7	3	5.54	6	5	4	0	2	1	0	2	1	0
7	4	5.83	6	5	4	0	2	1	0	3	2	1
8	1	5.26	7	6	5	0	3	2	1	0	0	0
8	2	5.28	7	6	5	0	3	2	1	1	0	0
8	3	5.56	7	6	5	0	3	2	1	2	1	0
8	4	5.84	7	6	5	0	3	2	1	3	2	1

LOSS RESERVING





- Semi-parametric approach works well for that
- Graph shows row and column factors from full regression vs. semi-parametric regression done in Stan
- Examples (and code) in my paper "Loss Reserving Using Estimation Methods Designed for Error Reduction" at the Variance Articles in Press site:
- https://www.variancejournal.org/articlespress/articles/Loss-Reserving-Venter.pdf

WRAP-UP

- Shrinking parameters in regressions also shrinks fitted values towards the mean and reduces the estimation and prediction variance, like shrinking towards the mean does in credibility theory
- There are random-effects and Bayesian forms of shrinkage – very similar
- Big advantages of Bayesian are:
  - Getting parameter uncertainty distributions
  - Easy and good penalized LL called loo
  - Can use fully Bayesian approach to simplify selecting degree of shrinkage
- Semi-parametric regression builds up customized curves across parameter types
- There is a fully Bayesian form of this from shrinking piecewise linear slope changes
- Not restricted to normal residuals – same thing works with GLM and even more general distributions of residuals. Also non-linear models. If you can write down the model's equations, you can put it in Stan, and then build curves across the parameters.
- All approaches give reduced error models for ratemaking and reserving