Claim Severities, Claim Relativities, and Age

Evidence from SOA Group Health Data

The paper of the same title is available for download at
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1412243

Chris Laws and Frank Schmid
NCCI, Inc.
Boca Raton, Florida

Annual Meeting
Casualty Actuarial Society
Boston, Mass.
November 15-18, 2009

Summary

- The influence of age on Group Health medical claims is analyzed for a large data set provided by the Society of Actuaries
- This data set comprises claims for the years 1997 through 1999, with a total claim count of about 4.3 million and total paid charges of approximately $7 billion
- Using partial linear models that allow for the influence of age to be nonlinear, it is shown how age affects…
  - Claim severities and…
  - Claim relativities (defined as the proportions of claimants by diagnostic category)

Objective

- Quantifying the effect of age on the severity of group health claim costs
  - There are two cost effects of age
    - Effect of age at given diagnosis
    - Effect of age on diagnosis
- By having the age effect quantified, it becomes possible to simulate claim severities and claim relativities for alternative projections of the future age composition of the working population (and its spouses and dependents)

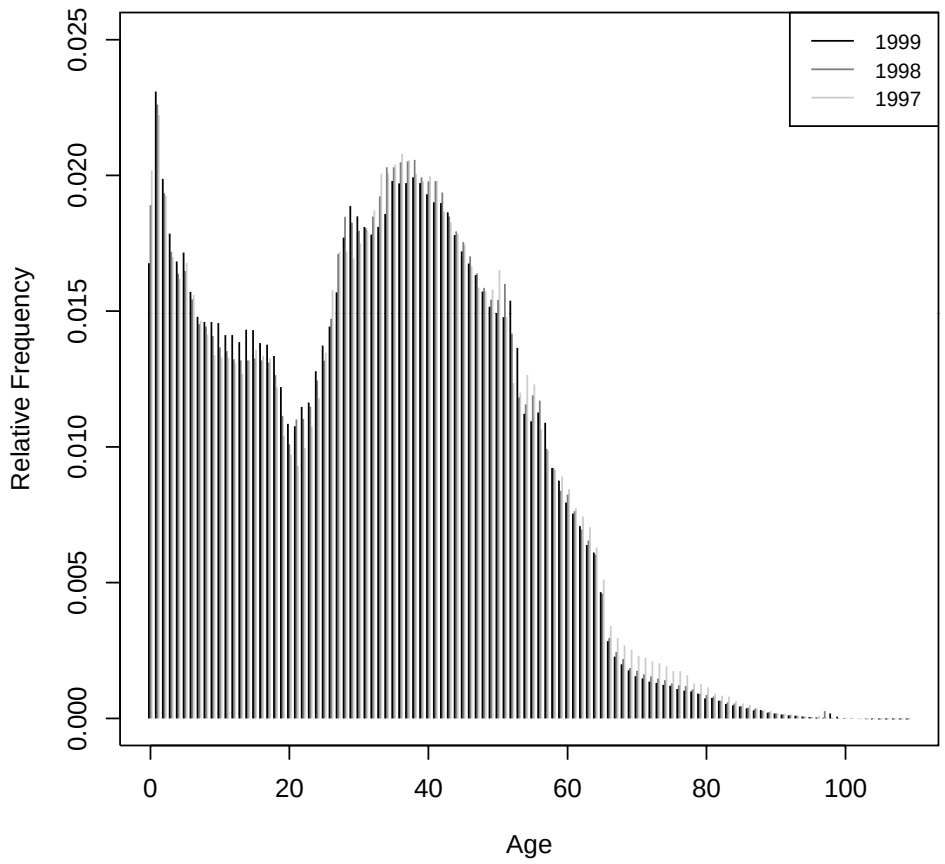
The Data Set
Data Items

- The SOA Group Health claims data base originates in a data call in which eleven insurers participated; the claims information of seven insurers was included in the final data set
- The resulting data base comprises the claim years 1997 through 1999, with a total of 4.3 million claims and $7.1 billion of paid charges (as incurred to the insurer)
- The severity of a claim is defined as total paid charges (as incurred to the insurer) for a single person (who may be an employee, spouse, or dependent) for the year
- Information is available on…
  - Enrollment status (employee, spouse, or dependent)
  - Gender
  - Type of care (PPO [preferred provider organization], or non-PPO)
  - Birth year
- Further, claims are classified by diagnostic category based on the highest subtotal of paid charges
- The data set is not longitudinal, which means that claimants cannot be followed over the years
- The data set is publicly available for download at http://www.soa.org/research/health/research-medical-large-claims-experience-study.aspx

Data source: SOA (Society of Actuaries, www.soa.org).
Note that claimants may not have been with the insurer for the entirety of the calendar year, for instance, where there are hires or layoffs.

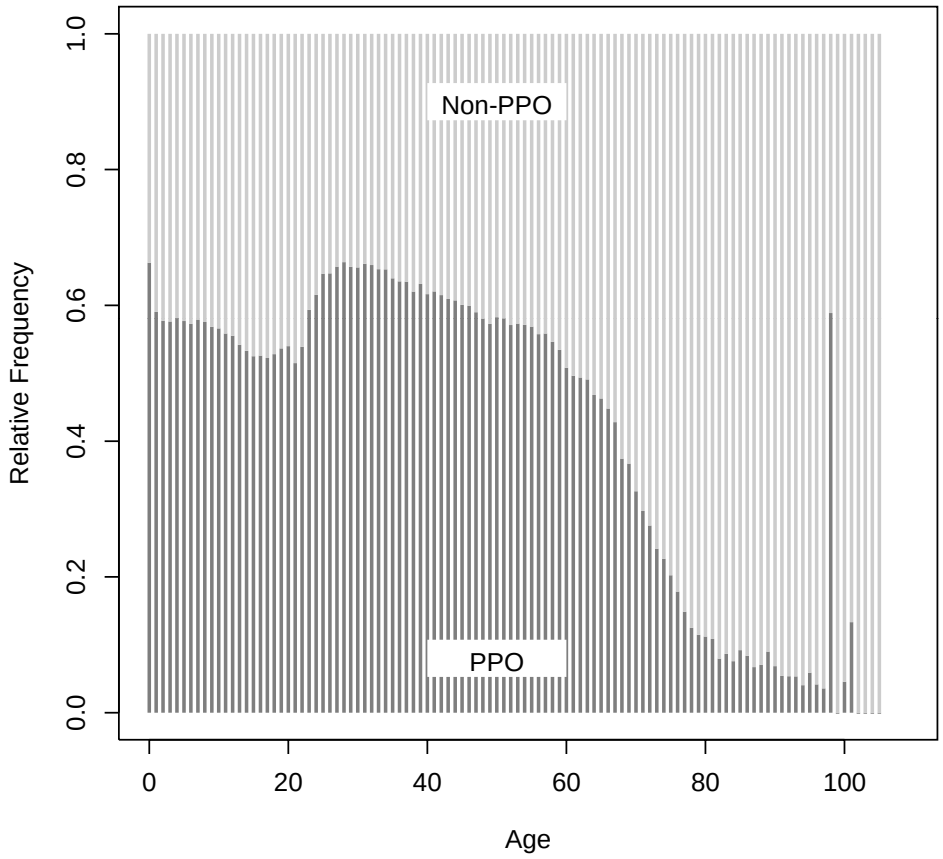
The Data Set

Frequency Distribution by Age



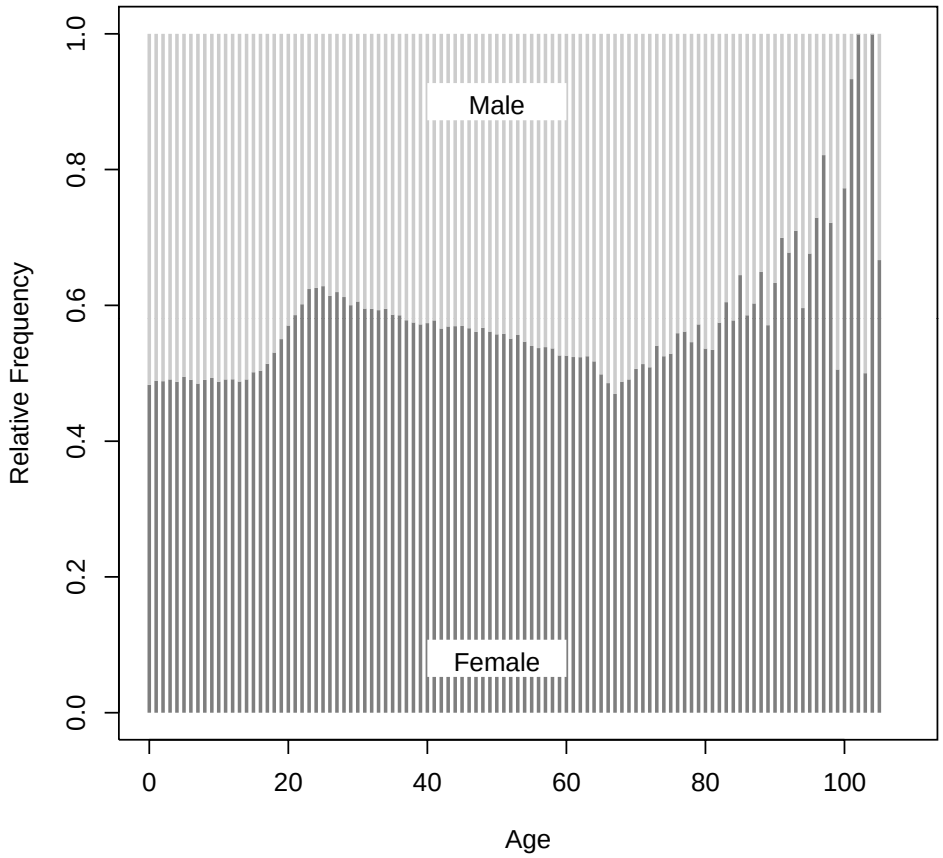
SOA Group Health medical claims, 1997-1999. The total number of claims equals 3,663,815.
Data source: SOA (Society of Actuaries, www.soa.org).

The Data Set
Frequency Distribution by Type of Care



SOA Group Health medical claims, 1999. The number of claims equals 1,378,244.
Data source: SOA (Society of Actuaries, www.soa.org).

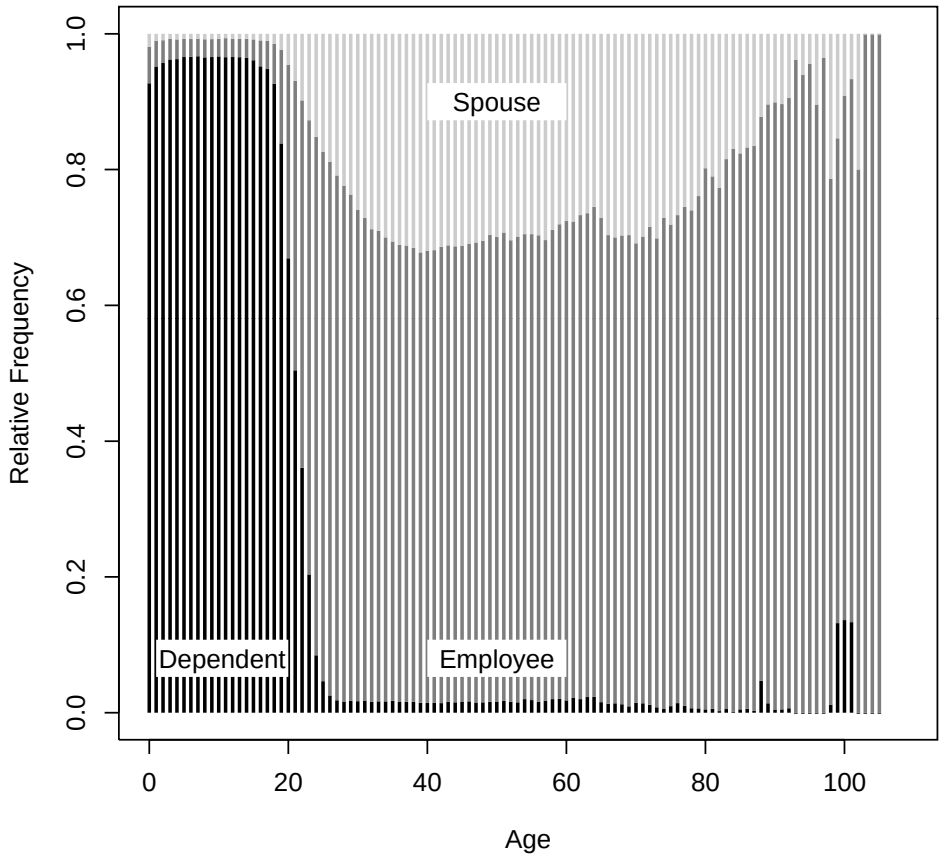
The Data Set
Frequency Distribution by Gender



SOA Group Health medical claims, 1999. The number of claims equals 1,378,244.
Data source: SOA (Society of Actuaries, www.soa.org).

The Data Set

Frequency Distribution by Enrollment Status



SOA Group Health medical claims, 1999. The number of claims equals 1,378,244.
Data source: SOA (Society of Actuaries, www.soa.org).

The Data Set
Categories of Diagnostic Codes

No.	ICD9 Range	Description	Number of Claims
1	001 through 139	Infectious & Parasitic Disease	95,329
2	140 through 239	Malignant Neoplasms	113,619
3	240 through 279	Endocrine & Metabolic Disorders	99,662
4	280 through 289	Blood Related Disorders	11,027
5	290 through 319	Mental Disorders, Drug, Alcohol	117,661
6	320 through 359	Nervous System	38,564
7	360 through 389	Sense Organs	205,727
8	390 through 459	Circulatory System	145,827
9	460 through 519	Respiratory System	433,814
10	520 through 579	Digestive System	143,046
11	580 through 629	Genitourinary System	242,882
12	630 through 679	Pregnancy & Childbirth	90,617
13	680 through 709	Skin Disorders	131,642
14	710 through 739	Skeleton & Muscle System	331,559
15	740 through 779	Congenital & Perinatal	23,012
16	780 through 799	Symptoms & Ill-Defined Conditions	422,664
17	800 through 999	Injury and Poisoning	424,270
18	V00 through V84	Health Status or Service	592,893

SOA Group Health medical claims, 1997-1999. The total number of claims equals 3,663,815. There are 18 categories of grouped ICD9 diagnostic codes.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities
Objective

- Gaining insights into the influence of age on claim severities by employing a partial linear quantile regression model, which allows for the claim severities to be…
  - Nonuniform in age
  - Nonlinear in age

Claim Severities
Review of Least Squares and ML

- Least squares (LS) and Maximum Likelihood (ML) regression quantify the effect of covariates on the mean of the outcome variable, thus assuming a uniform influence of the covariate on the quantiles of a distribution
  - LS estimates the regression coefficients by minimizing the sum of squared errors (between estimated and observed outcome)
  - ML estimates the regression coefficients by maximizing the likelihood of generating the observed outcome
- For the standard linear regression problem, LS and ML offer the same closed form for the regression coefficients
- Whereas LS makes no assumption regarding the distribution of the outcome variable, a common assumption in ML is that such distribution is conditionally normal

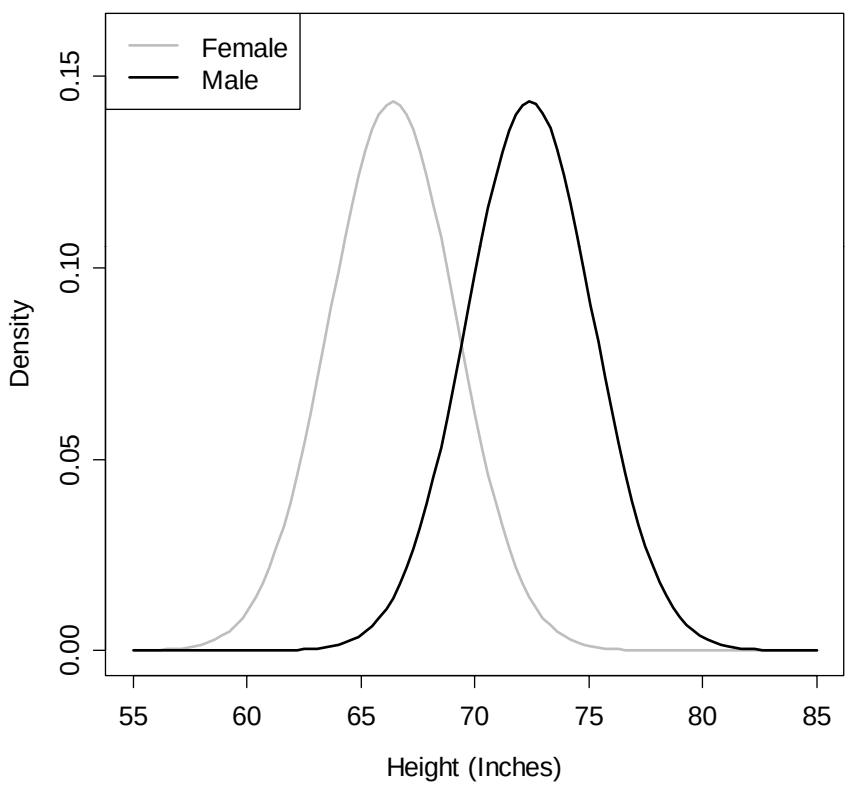
Claim Severities
Modeling the Conditional Mean (μ)

- For the purpose of simplicity, let us assume that the data are conditionally normal
  - In this case then, the standard regression equation reads

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$
$$\mu_i = \alpha + \beta \cdot x_i \ ,$$

where y is the outcome variable, μ is the conditional mean, x is the (single) covariate, and N indicates the normal distribution; (α, β) are the regression coefficients to be estimated

Claim Severities
Example of Shift in Mean

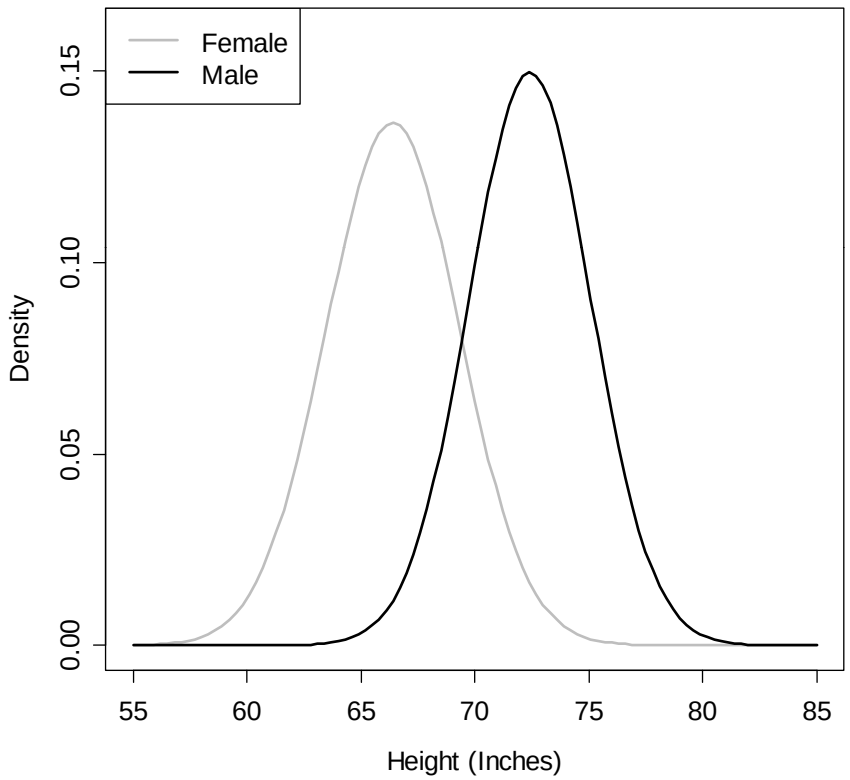


- Example: Height of students, regressed on gender
  - A normal distribution typically fits height well
  - The influence of height is uniform (that is, all the quantiles of the distribution shift by the same value)

Source: http://www.math.hope.edu/swanson/data/heights.txt.

Claim Severities

Example of Nonuniform Influence (Shift in Mean *and* Variance)



- Example, *cont.'d*
  - Gender not only affects the mean, but also the scale (as measured by the variance, for instance)
  - Males, in spite of being taller on average, exhibit a smaller variance (in this specific example)
  - Here, the effect of gender on the distribution of height is nonuniform (that is, the quantiles of the distribution do not all shift by the same value)

Source: http://www.math.hope.edu/swanson/data/heights.txt.

Claim Severities
Nonuniform Influence: Covariates May Affect Scale and Shape

- Generally, covariates may affect the distribution of an outcome variable with respect to its…
  - location (e.g., mean)
  - scale (e.g., variance), and
  - shape (e.g., skewness)
- There are maximum-likelihood approaches (inclusive of some Generalized Linear Models [GLMs]) that control for variation in both mean and variance
  - Yet, such models stay within known parametric distributions (or mixtures thereof)—by contrast, quantile regression makes no assumption regarding the distribution of the outcome variable

Claim Severities
A Brief History of Quantile Regression

- The first manuscript on an attempt to regress on the median was drafted by Rudjer Boscovich in 1760 (*)
- In 1978 Koenker and Basset (**) formulated the regression on the median as a linear programming problem, which can be solved using the Simplex algorithm (developed in 1947)
  - This approach can be generalized to regression on quantiles (other than the median) by means of asymmetric weighting of the errors
    - For instance, by weighting negative errors more highly than positive errors, the model regresses on quantiles smaller than the median

(*) Stigler, Stephen M., 1984, "Boscovich, Simpson, and a 1760 Manuscript Note on Fitting a Linear Relation," *Biometrika* **71**, 615-620.
(**) Koenker, Roger, and Gilbert Basset, Jr., (1978) "Regression Quantiles," *Econometrika* **46**, 33-50.

Claim Severities
Implementation of Quantile Regression

- An important catalyst for developing efficient algorithms and software for quantile regression was Gary Chamberlain's invited address to the 1990 World Congress of the Econometric Society, where he called for "going beyond models for the conditional mean" (*)
  - The R package quantreg (written by Roger Koenker) relies heavily on the interior point algorithm, various versions of which were developed in the 1990s
- Major advantages of quantile regression
  - Ability to regress on a location parameter other than the mean
  - It is possible to quantify the effect of covariates on location, scale, and shape of the distribution of the outcome variable
  - No assumption is necessary regarding the distribution of the outcome variable

(*) Koenker, Roger, and Kevin F. Hallock (2001) "Quantile Regression," *Journal of Economic Perspectives* **15**(4), 143-156.

Claim Severities
Outcome Variable, Covariates, and Reference Group

- Outcome variable
  - Total paid charges (as incurred to the insurer) for a single person (who may be an employee, spouse, or dependent) for the year
- The Covariates:
  - Age (measured by the difference between claim year and birth year)
  - Gender (male and female)
  - Type of managed care (PPO, non-PPO)
  - Diagnosis (ICD9 code, by highest subtotal of paid charges)—there are 18 categories
- The use of indicator (0/1) covariates (gender, type of managed care, and diagnosis) necessitates the choice of a reference group:
  - Male
  - Non-PPO
  - ICD9 codes V00-V84: Health Status or Service
    - V codes classify factors influencing health status and contact with health service—the circumstances are other than disease or injury

Claim Severities
The Model

- The model is of the semi-parametric (or, synonymously, partial linear) type:

$$y_i = \sum_{k=1}^{K} \beta_k^q \cdot x_{k,i} + f^q(z_i) \ ,$$

  where the matrix $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_K)$ comprises all covariates except age, and $\boldsymbol{\beta}^q$ is the vector of regression coefficients that gauges their influence on the location of quantile q of the log paid charges $\boldsymbol{y}$

- The smoother $f^q(z)$ represents the potentially nonlinear influence of age (z), evaluated at quantile q of the log paid charges $\boldsymbol{y}$

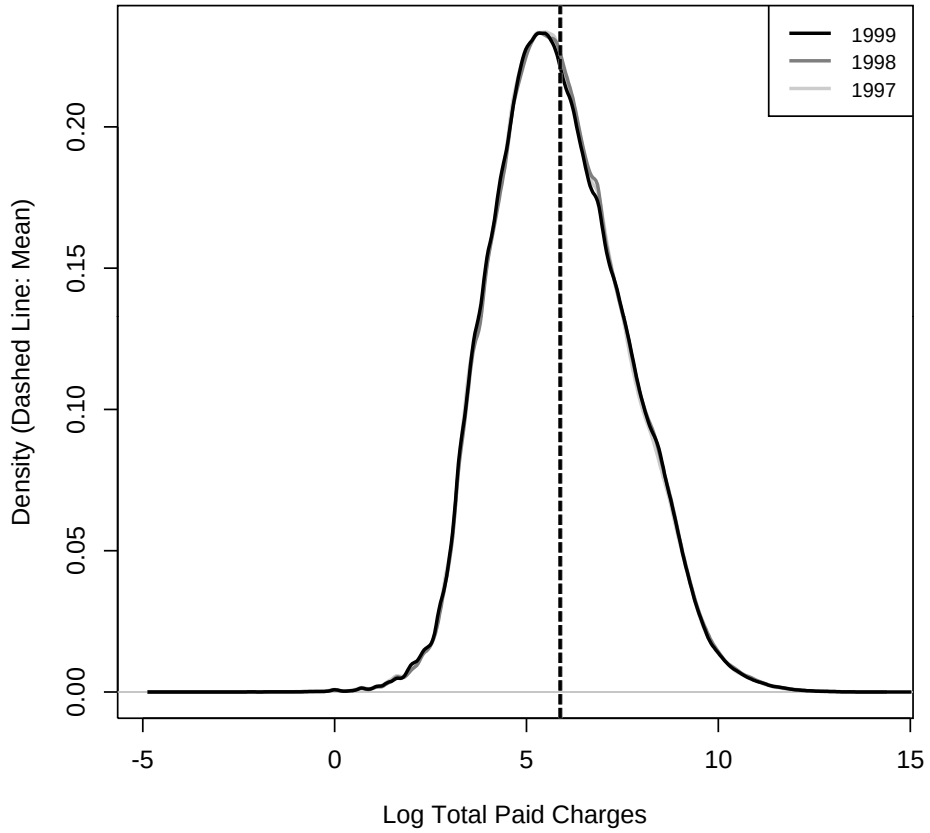
- The model is estimated using the R package `quantreg`

  - Total variation regularization is used for the smoother $f^q(z)$ (*)

(*) Koenker, Roger, Pin Ng, and Stephen Portnoy (1994) "Quantile Regression Smoothing," *Biometrika* **81**, 673-680.

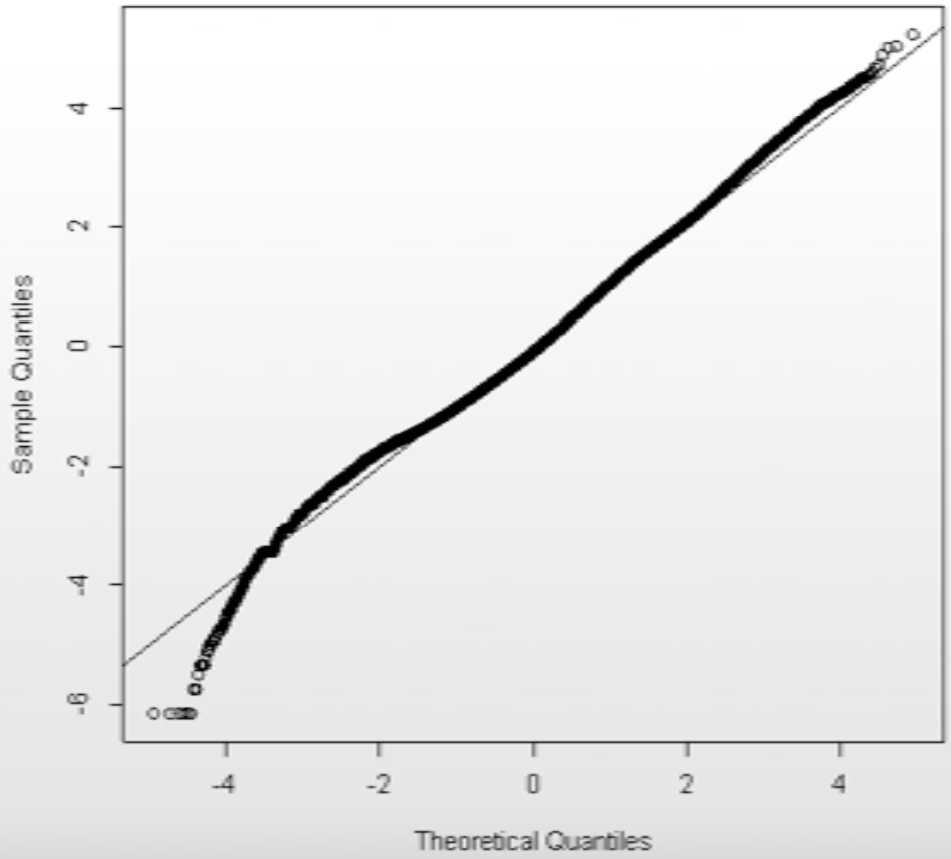
Claim Severities

Kernel Density of Log Paid Charges



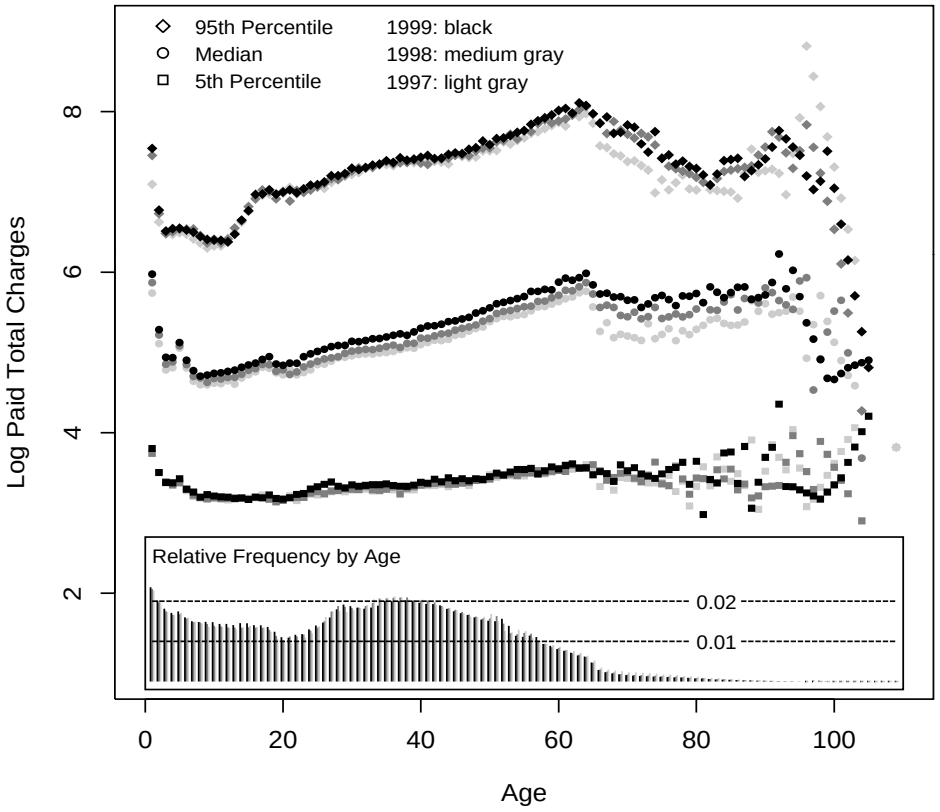
Logarithmic total paid charges of SOA Group Health medical claims, 1997-1999. The total number of claims equals 3,663,815. The dashed lines indicate the mean. There is positive skew, with the left tail being shorter than the right tail. Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities
Normal QQ-Plot of Log Paid Charges



Logarithmic total paid charges of SOA Group Health medical claims, 1999. The total number of claims equals 1,378,244. The QQ plot indicates that the left tail of the logarithmic empirical severity distribution is shorter than indicated by the normal distribution, while the right tail is only mildly thicker.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities
Quantile Smoothing of Age (All Claims)

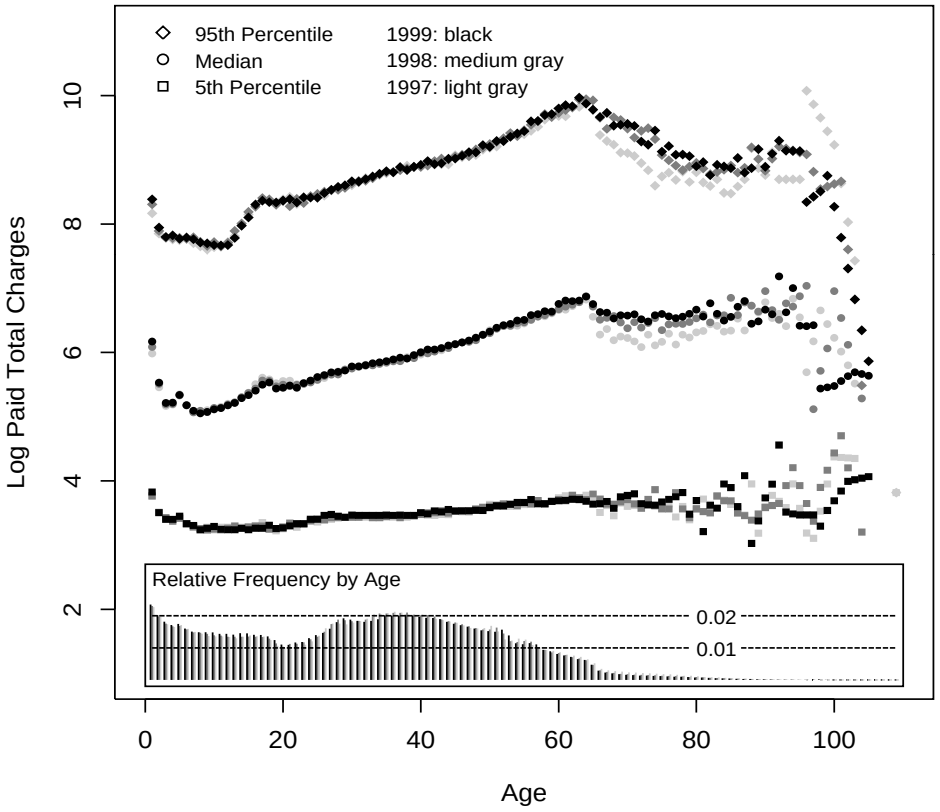


- For each quantile, paid total charges are almost perfectly log-linear in age from the early 20s through the early 60s
- An M estimator, applied to the estimated age effect within the age bracket [22,62], delivers growth rates (per year of age) of 0.75 percent (5th percentile), 2.39 percent (median), and 2.28 percent (95th percentile)
  - The differences in slopes (that is, in the growth rates) are indicative of a nonuniform influence of age on the paid charges
  - Note that these growth rates are net of the influence that age exerts on claim relativities by diagnostic category (net effect of age)

SOA Group Health medical claims, 1997-1999. The total number of claims equals 3,663,815. The partial linear model regresses logarithmic paid total charges on age (nonparametric component), gender, type of care, and ICD9 code, using the approach developed by Koenker, Ng, and Portnoy (1994). The intercept is added to the estimated value of the nonparametric component, which hosts the centered age covariate.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities

Quantile Smoothing of Age (All Claims, no ICD9 Codes)



- Again, for each quantile, paid total charges are almost perfectly log-linear in age from the early 20s through the early 60s
- An M estimator, applied to the estimated age effect within the age bracket [22,62], delivers growth rates (per year of age) of 0.82 percent (5th percentile), 3.13 percent (median), and 3.47 percent (95th percentile)
- Note that these growth rates incorporate the influence that age exerts on claim relativities by diagnostic category, as no indicator variables for ICD9 codes are included as covariates (gross effect of age)

SOA Group Health medical claims, 1997-1999. The number of claims equals 3,663,815. The partial linear model regresses logarithmic paid total charges on age (nonparametric component), gender, and type of care (but not ICD9 code), using the approach developed by Koenker, Ng, and Portnoy (1994). The intercept is added to the estimated value of the nonparametric component, which hosts the centered age covariate.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities
Effect of Age on Paid Charges (1999)

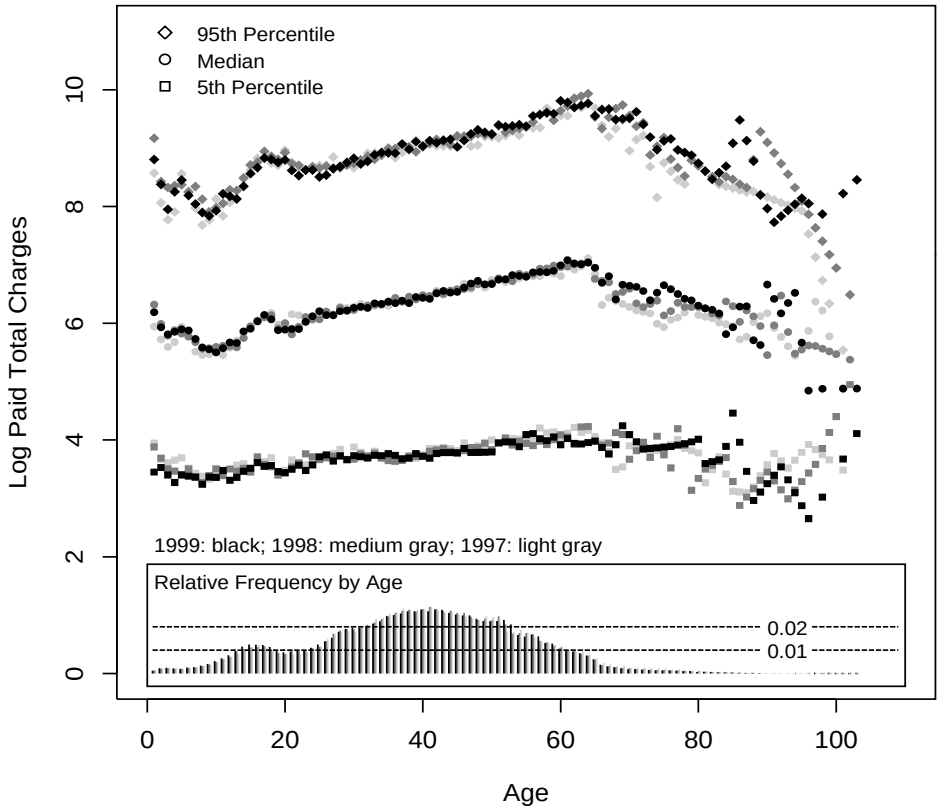
Age Effect	Percentage Increase of Paid Charges by Quantile, per Year of Age		
	0.05	**0.5**	**0.95**
Net (Controlling for ICD9 code)	0.75	2.39	2.28
Gross (No controlling for ICD9 code, except Pregnancy & Childbirth)	0.82	3.13	3.47

Based on the median, about three quarters of the gross age effect is due to higher paid charges at given diagnoses, while the remaining quarter is due to an age-related change in diagnoses

SOA Group Health medical claims, 1999. The number of claims equals 1,378,244. The partial linear model regresses logarithmic paid total charges on age (nonparametric component), gender, type of care, and ICD9 code (for the net age effect only), using the approach developed by Koenker, Ng, and Portnoy (1994). The effect of age is the slope of the nonparametric component over the age interval [22,62], as obtained by robust regression using an M estimator.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities

Quantile Smoothing of Age (Skeleton & Muscle System)

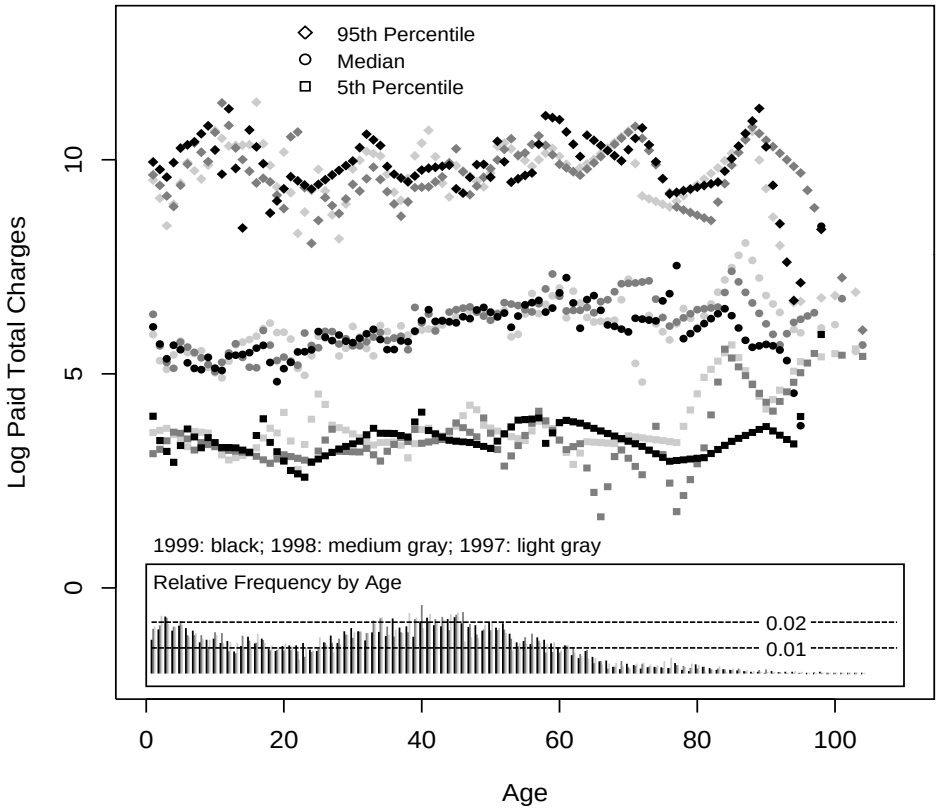


- Paid total charges are almost perfectly log-linear in age from the early 20s through the early 60s at each quantile

SOA Group Health medical claims, 1997-1999. The number of claims equals 331,559. The partial linear model regresses logarithmic paid total charges on age (nonparametric component), gender, and type of care, using the approach developed by Koenker, Ng, and Portnoy (1994). The intercept is added to the estimated value of the nonparametric component, which hosts the centered age covariate.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities

Quantile Smoothing of Age (Blood Related Disorders)

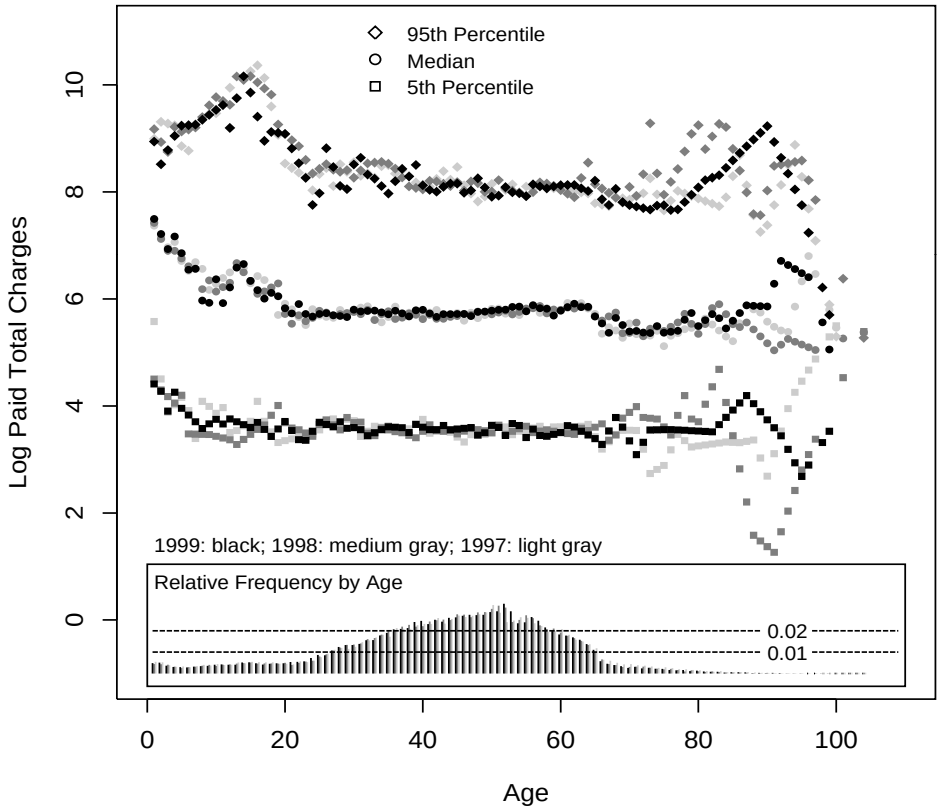


- Paid total charges vary systematically with age, but the relation is less smooth than for the category Skeleton & Muscle System
- The lack of smoothness is, at least in part, due to the comparatively small number of observations (of which there are 11,027, compared to 331,559 in the previously shown category Skeleton & Muscle System)

SOA Group Health medical claims, 1997-1999. The number of claims equals 11,027. The partial linear model regresses logarithmic paid total charges on age (nonparametric component), gender, and type of care, using the approach developed by Koenker, Ng, and Portnoy (1994). The intercept is added to the estimated value of the nonparametric component, which hosts the centered age covariate.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities

Quantile Smoothing of Age (Endocrine & Metabolic Disorders)



- Paid total charges vary systematically with age, and this relation is smooth (there are 99,662 observations)
- Most interestingly, the relation between log paid charges and age is negative at the 5th and 95th percentiles, but positive at the median

SOA Group Health medical claims, 1997-1999. The total number of claims equals 99,662. The partial linear model regresses logarithmic paid total charges on age (nonparametric component), gender, and type of care, using the approach developed by Koenker, Ng, and Portnoy (1994). The intercept is added to the estimated value of the nonparametric component, which hosts the centered age covariate.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Severities

Effect of Age on Paid Charges by Diagnostic Code

Diagnosis Category	Percentage Increase of Paid Charges by Quantile, per Year of Age		
	0.05	0.5	0.95
Infectious & Parasitic Disease	0.36	3.17	6.66
Malignant Neoplasms	2.58	5.06	5.17
Endocrine & Metabolic Disorders	-0.11	0.29	-0.76
Blood Related Disorders	1.82	2.96	1.63
Mental Disorders, Drug, Alcohol	0.91	2.17	1.63
Nervous System	1.50	3.53	2.79
Sense Organs	0.89	2.74	1.58
Circulatory System	0.53	2.33	4.95
Respiratory System	0.33	3.12	4.65
Digestive System	2.40	3.61	1.80
Genitourinary System	0.74	1.76	1.38
Pregnancy & Childbirth	0.81	1.12	2.28
Skin Disorders	0.07	2.03	2.77
Skeleton & Muscle System	1.06	2.43	2.91
Congenital & Perinatal	0.43	2.39	0.53
Symptoms & Ill-Defined Conditions	1.50	3.23	2.56
Injury and Poisoning	0.70	2.48	2.54
Health Status or Service	-0.29	0.97	---

Group Health medical claims, 1999. The number of claims equals 1,378,244. The partial linear model regresses by diagnosis logarithmic paid total charges on age (nonparametric component), gender (except for Pregnancy & Childbirth), and type of care, using the approach developed by Koenker, Ng, and Portnoy (1994). The effect of age is the slope of the nonparametric component over the age interval [22,62] (or [20,40] for Pregnancy & Childbirth), as obtained by robust regression using an M estimator. No value is provided for the 95th quantile of Health Status or Service, the age effect of which does not display the familiar log-linearity.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Relativities
The Multinomial Logit Model

- Claim relativities are modeled using a semi-parametric multinomial logit specification:

$$\boldsymbol{y}_i = y_{i,1\ldots J} = Multinomial\left(p_{i,1\ldots J}, N_i\right) \quad , \quad N_i = \sum_{j=1}^{J} y_{i,j} \ , \ i = 1,\ldots,I$$

$$p_{i,j} = \frac{\exp\left(\boldsymbol{x}_i \cdot \boldsymbol{\beta}_j + f_j(z_i)\right)}{1 + \sum_{j=1}^{J-1} \exp\left(\boldsymbol{x}_i \cdot \boldsymbol{\beta}_j + f_j(z_i)\right)} \quad , \qquad j = 1,\ldots,J-1$$

$$p_{i,J} = \frac{1}{1 + \sum_{j=1}^{J-1} \exp\left(\boldsymbol{x}_i \cdot \boldsymbol{\beta}_j + f_j(z_i)\right)} \quad ,$$

where $\boldsymbol{y}_i$ is a count vector for claimants of type i. This vector has 18 elements (or 17, when Childbirth & Pregnancy is excluded). Element $y_{i,j}$ of vector $\boldsymbol{y}_i$ represents the number of claimants of type i that fall into diagnostic category j. The probability with which a claimant of type i belongs to diagnostic category j is governed by the parameter vector $\boldsymbol{p}_i$, which is to be estimated

The multinomial model is estimated separately for male and female claimants. The number of types of claimants equals the number of unique age values of PPO claimants plus the number of unique age values of non-PPO claimants. Grouping the claimant level data into claimant types considerably reduces the computational task of estimating a multinomial logit model for several hundred thousand claimants by means of MCMC.
Data source: SOA (Society of Actuaries, www.soa.org).

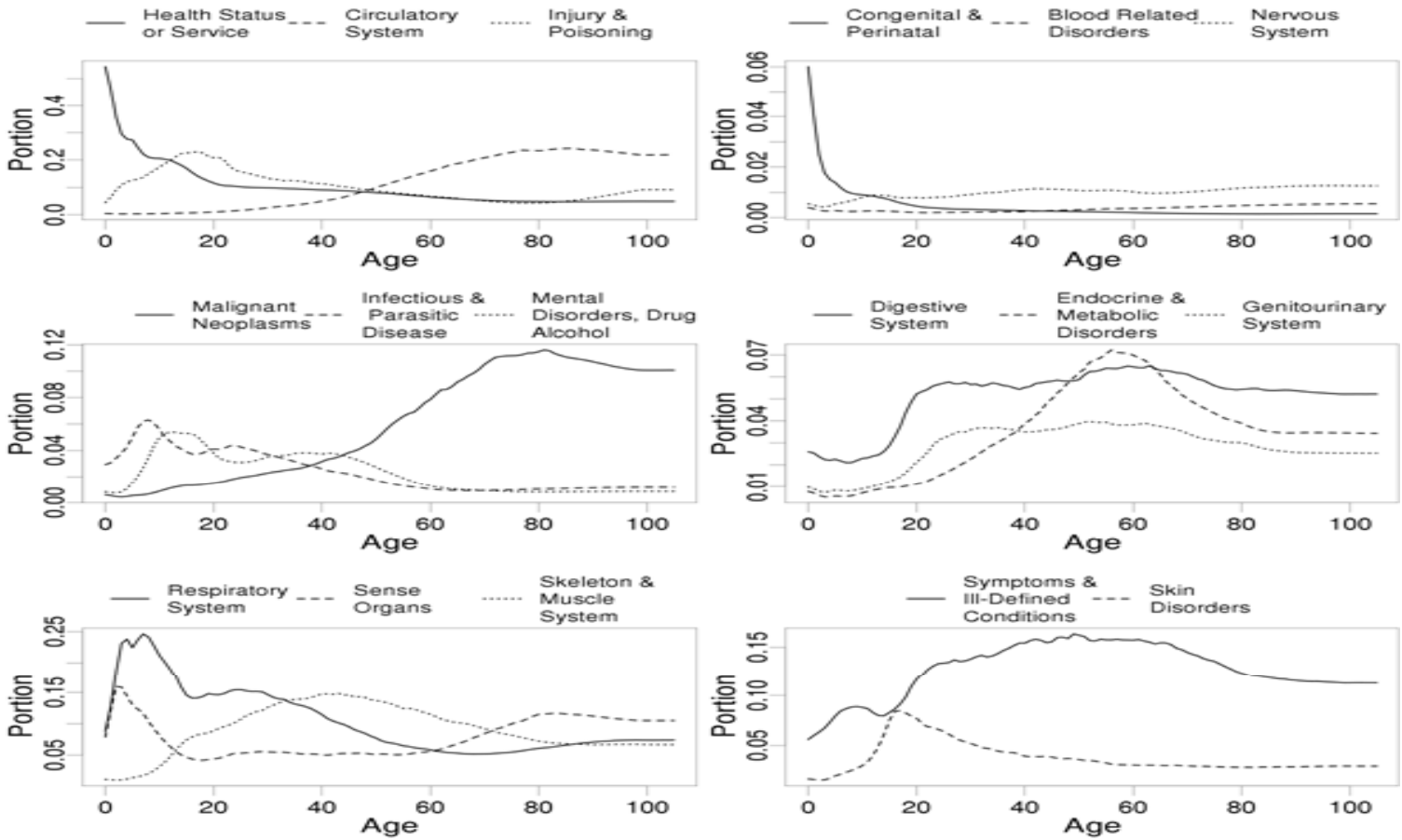
Claim Relativities
The Multinomial Logit Model, cont'd

- The multinomial model is estimated using Markov Chain Monte Carlo simulation (MCMC)
- The smoother $f(z)$ is a first-order random walk specification (*)
- Grouping the claimant level data into *I* claimant types considerably reduces the computational task of estimating this model for several hundred thousand claimants by means of MCMC

(*) Fahrmeier, Ludwig, and Stefan Lang (2001) "Bayesian Inference for Generalized Additive Mixed Models based on Markov Random Field Priors," *Journal of the Royal Statistical Society*, Series C, **50**, 201-220

Claim Relativities

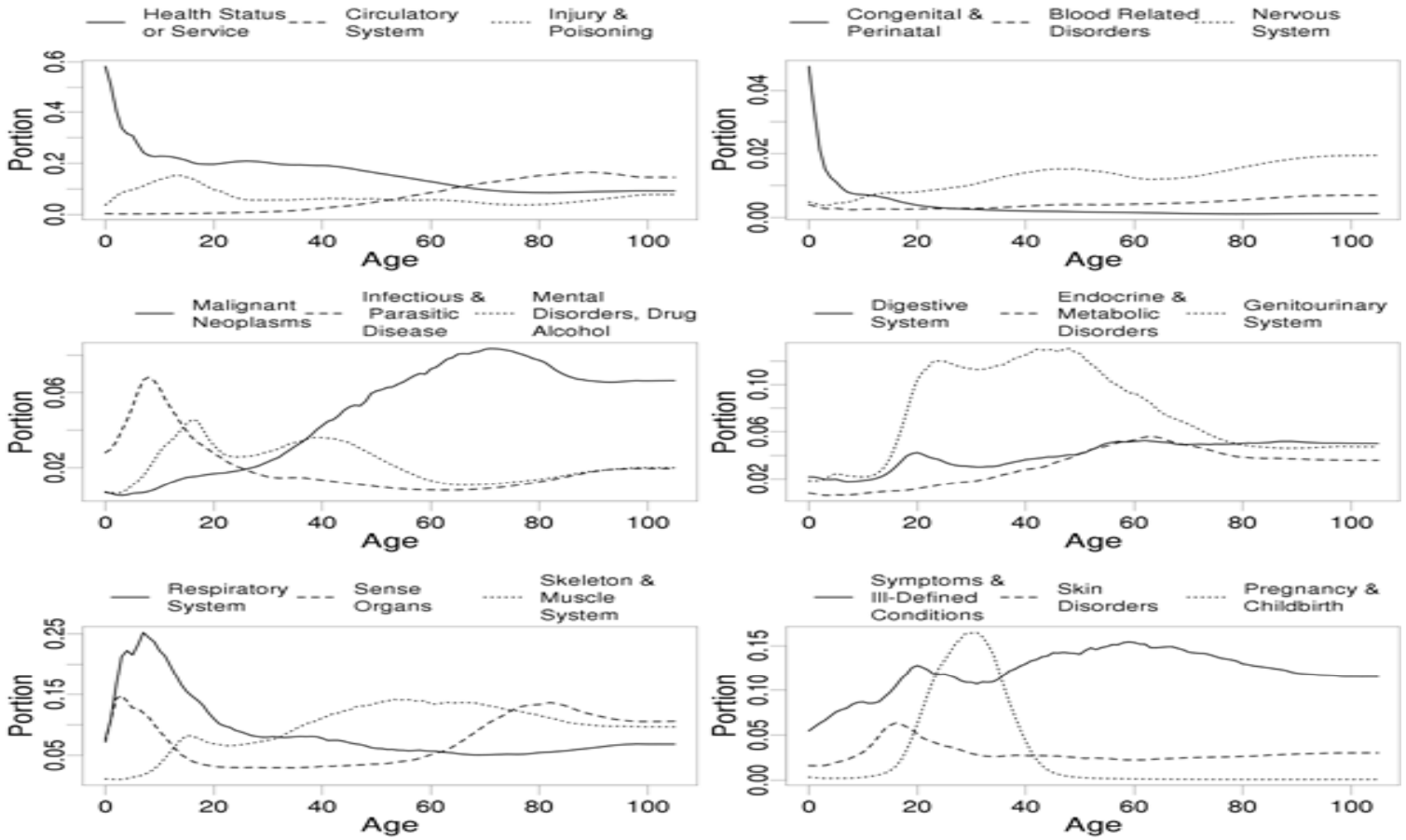
Multinomial Logit Model, Male Claimants



SOA Group Health medical claims, 1999. The total number of claims equals 620,690. The partial linear multinomial logit model regresses the diagnostic category on age (nonparametric component) and type of care, using a random walk smoother in the non parametric component. The intercept is added to the estimated value of the nonparametric component, which hosts the age covariate. The reference group is non-PPO.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Relativities

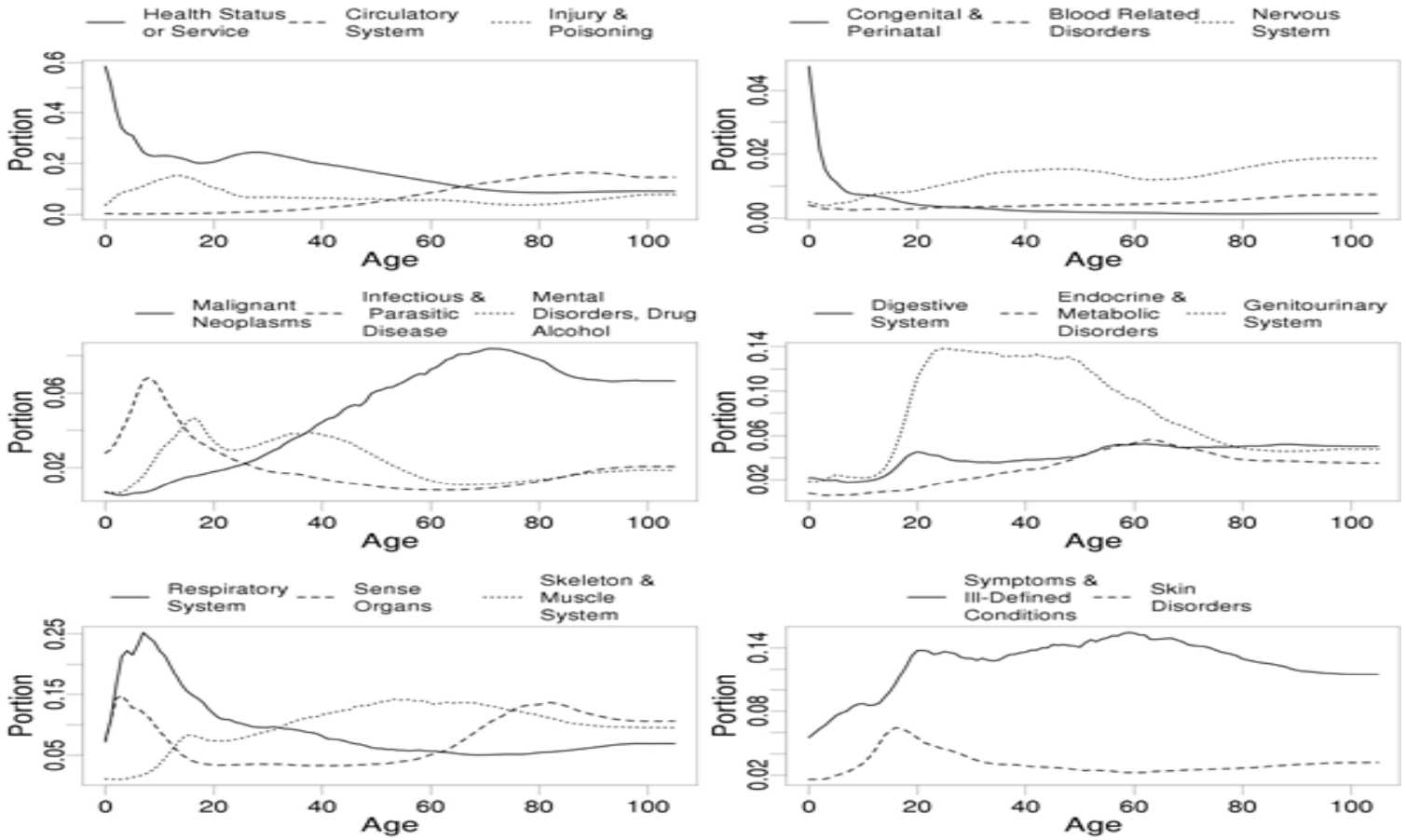
Multinomial Logit Model, Female Claimants



SOA Group Health medical claims, 1999. The total number of claims equals 757,554. The partial linear multinomial logit model regresses the diagnostic category on age (nonparametric component) and type of care, using a random walk smoother in the non parametric component. The intercept is added to the estimated value of the nonparametric component, which hosts the age covariate. The reference group is non-PPO.
Data source: SOA (Society of Actuaries, www.soa.org).

Claim Relativities

Multinomial Logit Model, Female Claimants Excluding Pregnancy & Childbirth



SOA Group Health medical claims, 1999. SOA Group Health medical claims, 1999. The total number of claims equals 725,077. The partial linear multinomial logit model regresses the diagnostic category on age (nonparametric component) and type of care, using a random walk smoother in the non parametric component. The intercept is added to the estimated value of the nonparametric component, which hosts the age covariate. The reference group is non-PPO..
Data source: SOA (Society of Actuaries, www.soa.org).

Conclusion
Claim Severities

- Using a partial linear model that allows for a nonlinear influence of age in the nonparametric component, we found that the relation between Group Health claim costs (in logarithmic terms) and age is linear from the early twenties to the early sixties, but nonlinear outside this age bracket
  - Most interestingly, this log-linear relation between claim severity and age at the level of the individual claimant holds up for nearly every diagnostic category
- Further, we were able to show that the percentage impact of age on paid charges is greater in the right tail of the severity distribution than at the median or in the left tail

Conclusion
Claim Relativities

- We also studied the effect of age on claim relativities, which are defined as the proportions of claimants that fall into the various diagnostic categories; the diagnostic category of a claimant is determined by the highest subtotal of paid charges for the year
  - The claim relativities were analyzed using a partial linear multinomial logit model, where the nonparametric component again comprises the influence of age
- We found that, similar to the intervention rates studied by Polder et al. (*), claim relativities are highly nonlinear in age

(*) Polder, Johan J., Luc Bonneux, Willem Jan Meerding, and Paul J. van der Maas (2002) "Age-Speficic Increases in Health Care Costs," *European Journal of Public Health* **12**, 57-62.

Appendix
Data Cleansing

- Prior to analyzing the data set, we purge it of claims that appear to present mis-coded or sparse information. For instance, of the original 4,775,214 claims, 144,089 claimants (or three percent of the original 4,775,214 records) exhibit a birth year greater than the claim year, thus implying negative age. Further, claims with unknown diagnosis category (631,836 records) are dropped, as well as claims with an enrollment status other than employee, spouse, or dependent (549 records). In addition, there are 7,386 records of male claimants with diagnostic code Childbirth & Pregnancy; these claims are discarded as well. Finally, for reasons of sparseness, we discard nine claims (possibly comprising three claimants for three years) with a claimant's age of 127 or higher, thus leaving the data set with a maximum age of 104 years.

Data source: SOA (Society of Actuaries, www.soa.org).