Predictive Modeling of Insurance Company Operations

Edward W. (Jed) Frees

University of Wisconsin – Madison

May, 2013

Outline

An Actuary Is ...

Predictive analytics

- is an area of statistical analysis that deals with
  - extracting information from data and
  - using it to predict future trends and behavior patterns.
- relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes.
- is used in financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields.

Predictive Modeling and Insurance

- Initial Underwriting
  - Offer right price for the right risk
  - Avoid adverse selection
- Renewal Underwriting/Portfolio Management
  - Retain profitable customers longer
- Claims Management
  - Manage claims costs
  - Detect and prevent claims fraud
  - Understand excess layers for reinsurance and retention
- Reserving
  - Provide management with an appropriate estimate of future obligations
  - Quantify the uncertainty of the estimates

Business Analytics and Insurance

- Sales and Marketing
  - Predict customer behavior and needs, anticipate customer reactions to promotions
  - Reduce acquisition costs (direct mail, discount programs)
- Compensation Analysis
  - Incent and reward employee/agent behavior appropriately
- Productivity Analysis
  - Analyze production of employees, other units of business
  - Seek to optimize production
- Financial Forecasting

Predictive Models

Here are some useful skills/topics

- Two-Part? For example, loss or no loss
- Loss distributions are typically skewed and heavy-tailed

Predictive Models

Here are some useful skills/topics

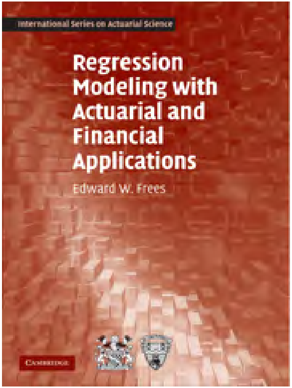
- Two-Part? For example, loss or no loss
- Loss distributions are typically skewed and heavy-tailed
- Censored?
  - Losses censored by amounts through deductibles or policy limits
  - Loss censored by time, e.g., claim triangles
- Insurance data typically has lots of explanatory variables. Lots.

Predictive Modeling and Statistics

- I think about predictive modeling as a subset of business analytics, although many use the terms interchangeably
- For some, predictive modeling means advanced data-mining tools as per Hastie, Tibshirani and Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.*
  - These tools include neural networks, classification trees, nonparametric regression and so forth
- Others think about the traditional triad of statistical inference:
  - Estimation
  - Hypothesis Testing
  - Prediction
- I fall in this latter camp

Some Advertising

- See my latest book: Frees (2010), *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press.
- As indicated by the title, the focus here is on regression



Thanks to ...

Collaborators:

- Emiliano Valdez, Katrien Antonio, Margie Rosenberg
- Peng Shi, Yunjie (Winnie) Sun
- Glenn Meyers, A. David Cummings
- Xipei Yang, Zhengjun Zhang, Xiaoli Jin, Xiao (Joyce) Lin

Motivating Two-Part Models

- Insurance and healthcare data often feature a large proportion of zeros, where zero values can represent:
  - Individual's lack of utilization
  - No expenditure (e.g., no claim)
  - Non-participation in a program

Motivating Two-Part Models

- Insurance and healthcare data often feature a large proportion of zeros, where zero values can represent:
  - Individual's lack of utilization
  - No expenditure (e.g., no claim)
  - Non-participation in a program
- How to model zero expenditures?
  - Ignore their existence
  - Throw them out and condition that usage is greater than zero
  - Do something else

Two-Part Models

- Economists use the term 'two-part models' (First part = whether zero, or $>$ 0; Second part = Amount)
- Actuaries refer to these as frequency and severity models and introduced in Bowers et al. (Chapter 2)
  - Let r_i = 1, if claim, 0 otherwise
  - y_i = amount of the claim.
  - $(Claim\ recorded)_i = r_i \times y_i$
- Two-part models include covariates in each part.

Two-Part Models

- Economists use the term ’two-part models’ (First part = whether zero, or $>$ 0; Second part = Amount)
- Actuaries refer to these as frequency and severity models and introduced in Bowers et al. (Chapter 2)
  - Let r_i = 1, if claim, 0 otherwise
  - y_i = amount of the claim.
  - $(Claim\ recorded)_i = r_i \times y_i$
- Two-part models include covariates in each part.
- I will use data from the Medical Expenditure Panel Survey (MEPS) to illustrate a few ideas
  - y = Medical Expenditure, many x’s to explain/predict

Inpatient Expenditures Summary Statistics

Category	Variable	Description	Percent of data	Average Expend	Percent Positive Expend	Average of Pos Expend
Demography	AGE	Age in years between 18 to 65 (mean: 39.0)				
	GENDER	1 if female	52.7	0.91	10.7	8.53
	GENDER	1 if male	47.3	0.40	4.7	8.66
Ethnicity	ASIAN	1 if Asian	4.3	0.37	4.7	7.98
	BLACK	1 if Black	14.8	0.90	10.5	8.60
	NATIVE	1 if Native	1.1	1.06	13.6	7.79
	WHITE	Reference level	79.9	0.64	7.5	8.59
Region	NORTHEAST	1 if Northeast	14.3	0.83	10.1	8.17
	MIDWEST	1 if Midwest	19.7	0.76	8.7	8.79
	SOUTH	1 if South	38.2	0.72	8.4	8.65
	WEST	Reference level	27.9	0.46	5.4	8.51
Education	COLLEGE	1 if college or higher degree	27.2	0.58	6.8	8.50
	HIGHSCHOOL	1 if high school degree	43.3	0.67	7.9	8.54
	Reference level is lower than high school degree		29.5	0.76	8.8	8.64
Self-rated physical health	POOR	1 if poor	3.8	3.26	36.0	9.07
	FAIR	1 if fair	9.9	0.66	8.1	8.12
	GOOD	1 if good	29.9	0.70	8.2	8.56
	VGOOD	1 if very good	31.1	0.54	6.3	8.64
	Reference level is excellent health		25.4	0.42	5.1	8.22
Self-rated mental health	MNHPOOR	1 if poor or fair	7.5	1.45	16.8	8.67
		0 if good to excellent mental health	92.5	0.61	7.1	8.55
Any activity limitation	ANYLIMIT	1 if any functional or activity limitation	22.3	1.29	14.6	8.85
		0 if otherwise	77.7	0.50	5.9	8.36
Income compared to poverty line	HINCOME	1 if high income	31.6	0.47	5.4	8.73
	MINCOME	1 if middle income	29.9	0.61	7.0	8.75
	LINCOME	1 if low income	15.8	0.73	8.3	8.87
	NPOOR	1 if near poor	5.8	0.78	9.5	8.19
	Reference level is poor/negative		17.0	1.06	13.0	8.18
Insurance coverage	INSURE	1 if covered by public or private health insurance in any month of 2003	77.8	0.80	9.2	8.68
		0 if have not health insurance in 2003	22.3	0.23	3.1	7.43
Total			100.0	0.67	7.9	8.32

MEPS Data: Random sample of 2,000 individuals aged 18 - 64 from first panel in 2003.

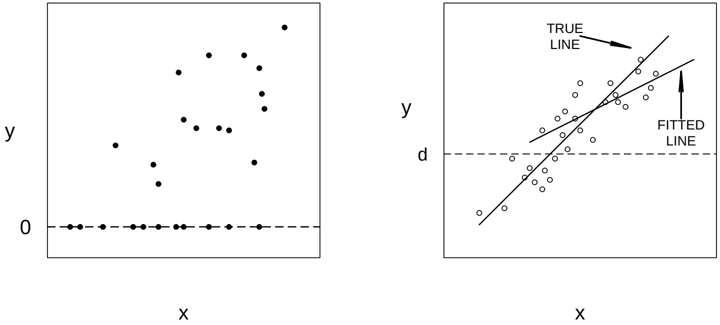
Inpatient Expenditures Summary Statistics

Category	Variable	Description	Percent of data	Average Expend	Percent Positive Expend	Average of Pos Expend
Demography	AGE	Age in years between 18 to 65 (mean: 39.0)				
	GENDER	1 if female	52.7	0.91	10.7	8.53
	GENDER	1 if male	47.3	0.40	4.7	8.66
Total			100.0	0.67	7.9	8.32

MEPS Data: Random sample of 2,000 individuals aged 18 - 64 from first panel in 2003.

Bias Due to Limited Dependent Variables

- Either excluding or ignoring zeros induces a bias
  - Left-hand panel: When individuals do have health expenditures, they are recorded as $y = 0$ expenditures. (Censored)
  - Right-hand panel: If the responses below the horizontal line at $y = d$ are omitted, then the fitted regression line is very different from the true regression line. (Truncated)



How do we estimate model parameters?

Use maximum likelihood. Standard calculations show $\ln L$ as:

$$
\begin{aligned}
\ln L &= \sum_{i:y_i=d_i} \ln\left(1-\Phi\left(\frac{\mathbf{x}_i'\beta - d_i}{\sigma}\right)\right) \\
&\quad -\frac{1}{2}\sum_{i:y_i>d_i}\left\{\ln 2\pi\sigma^2 + \frac{(y_i-(\mathbf{x}_i'\beta - d_i))^2}{\sigma^2}\right\}
\end{aligned}
$$

where $\{i : y_i = d_i\}$ = Sum of censored observations
and $\{i : y_i > d_i\}$ = Sum over non-censored observations.

Definition of Two-Part Model

1. Use a binary regression model with r_i as the dependent variable and $\mathbf{x}_{1i}$ as the set of explanatory variables.
   - Denote the corresponding set of regression coefficients as β_1.
   - Typical models include the linear probability, logit and probit models.
2. Conditional on $r_i = 1$, specify a regression model with y_i as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables.
   - Denote the corresponding set of regression coefficients as β_2.
   - Typical models include the linear regression and gamma regression models.

	Full Model				Reduced Model			
	Frequency		Severity		Frequency		Severity	
Effect	Estimate	*t*-ratio	Estimate	*t*-ratio	Estimate	*t*-ratio	Estimate	*t*-ratio
Intercept	-2.263	-10.015	6.828	13.336	-2.281	-11.432	6.879	14.403
AGE	-0.001	-0.154	0.012	1.368			0.020	2.437
GENDER	0.395	4.176	-0.104	-0.469	0.395	4.178	-0.102	-0.461
ASIAN	-0.108	-0.429	-0.397	-0.641	-0.108	-0.427	-0.159	-0.259
BLACK	0.008	0.062	0.088	0.362	0.009	0.073	0.017	0.072
NATIVE	0.284	0.778	-0.639	-0.905	0.285	0.780	-1.042	-1.501
NORTHEAST	0.283	1.958	-0.649	-2.035	0.281	1.950	-0.778	-2.422
MIDWEST	0.239	1.765	0.016	0.052	0.237	1.754	-0.005	-0.016
SOUTH	0.132	1.099	-0.078	-0.294	0.130	1.085	-0.022	-0.081
COLLEGE	0.048	0.356	-0.597	-2.066	0.049	0.362	-0.470	-1.743
HIGHSCHOOL	0.002	0.017	-0.415	-1.745	0.003	0.030	-0.256	-1.134
POOR	0.955	4.576	0.597	1.594	0.939	4.805		
FAIR	0.087	0.486	-0.211	-0.527	0.079	0.450		
GOOD	0.184	1.422	0.145	0.502	0.182	1.412		
VGOOD	0.095	0.736	0.373	1.233	0.094	0.728		
MNHPOOR	-0.027	-0.164	-0.176	-0.579			-0.177	-0.640
ANYLIMIT	0.318	2.941	0.235	0.981	0.311	3.022	0.245	1.052
HINCOME	-0.468	-3.131	0.490	1.531	-0.470	-3.224		
MINCOME	-0.314	-2.318	0.472	1.654	-0.314	-2.345		
LINCOME	-0.241	-1.626	0.550	1.812	-0.241	-1.633		
NPOOR	-0.145	-0.716	0.067	0.161	-0.146	-0.721		
INSURE	0.580	4.154	1.293	3.944	0.579	4.147	1.397	4.195
Scale σ^2			1.249				1.333	

The outcome of interest is $y = \begin{cases} 0 & r = 0 \\ y^* & r = 1 \end{cases}$

- r indicates if a claim has occurred and,
- conditional on claim occurrence ($r = 1$), y^* is the claim amount.

Part 1. The distribution of r can be written as $\mathrm{F}_r(\theta_r)$, where the parameter vector depends on explanatory variables $\theta_r = \theta_r(\mathbf{x})$.
Part 2. Similarly, the distribution of y^* can be written as $\mathrm{F}_y(\theta_y)$, where $\theta_y = \theta_y(\mathbf{x})$.

- When θ_r and θ_y are functionally independent, we can optimize each part in isolation of one another and thus, treat the likelihood process in “two parts.”

Alternatives to the Two-Part Model

- Tobit model. A related model used extensively in econometrics, where $y = \max(0, y^*)$. This is a censored regression model.
  - The tobit regression model typically assumes normality. In contrast, the two-part model retains flexibility in the specification of the amount distribution.

Alternatives to the Two-Part Model

- Tobit model. A related model used extensively in econometrics, where $y = \max(0, y^*)$. This is a censored regression model.
  - The tobit regression model typically assumes normality. In contrast, the two-part model retains flexibility in the specification of the amount distribution.
- Tweedie GLM. Compared to the two-part model, a strength of the Tweedie approach is that both parts are estimated simultaneously; this means fewer parameters, making the variable selection process simpler.
  - The Tweedie distribution is a Poisson sum of gamma random variables.
  - Thus, it has a mass at zero as well as a continuous component.
  - It is used to model "pure premiums," where the zeros correspond to no claims and the positive part is used for the claim amount.

Aggregate Loss Model

- For an aggregate loss model, we observe $y=(N,S_N)$ and $\mathbf{x}$
  - N describes the number of claims
  - S_N is the aggregate claim amount.
  - As with the two-part model, we separate the count (N) and severity portions (S_N).
- Alternatively, we may observe $(N,y_1^*,\ldots,y_N^*,\mathbf{x})$.
  - y_j^* describes the claim amount for each event/episode.
  - $S_N=y_1^*+\cdots+y_N^*$ is the aggregate claim amount.

Predictive Models

Here are some skills/topics useful in predictive modeling

- Two-Part? For example, loss or no loss
- Loss distributions are typically skewed and heavy-tailed
- Censored?
  - Losses censored by amounts through deductibles or policy limits
  - Loss censored by time, e.g., claim triangles
- Insurance data typically has lots of explanatory variables. Lots.

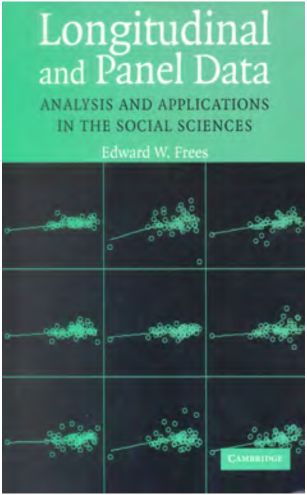
Predictive Models

Here are some skills/topics useful in predictive modeling

- Two-Part? For example, loss or no loss
- Loss distributions are typically skewed and heavy-tailed
- Censored?
  - Losses censored by amounts through deductibles or policy limits
  - Loss censored by time, e.g., claim triangles
- Insurance data typically has lots of explanatory variables. Lots.
- Multivariate responses, e.g., types of coverages, perils, bundling of insurances
- Longitudinal (panel)? Are you following the contract over time?
- Losses credible? we often wish to incorporate external knowledge into our analysis

Advertising 2

See, for example, my 2004 book.



Multivariate Regression

In multivariate analysis, there are several outcomes of interest (multivariate), $\mathbf{y}$

With regression, there are several variables available to explain/predict these outcomes, $\mathbf{x}$

Multivariate regression provides the foundation for several statistical methodologies.

- Structural Equations Modeling (SEM)
- Longitudinal Data Modeling
- Hierarchical Linear Modeling

Now suppose the outcome of interest is $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$.

- Use the notation $\mathrm{F}_j(\theta_j)$ for the distribution function of y_j, $j = 1, \ldots, p$.

Multivariate Regression

Now suppose the outcome of interest is $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$.

- Use the notation $\mathrm{F}_j(\theta_j)$ for the distribution function of y_j, $j = 1, \ldots, p$.
- The joint distribution function can be expressed using a copula C as

$$F = C(F_1, \ldots, F_p).$$

- The set of parameters is
  - $\theta = \{\theta_1, \ldots, \theta_p, \alpha\}$,
  - where α is the set of parameters associated with the copula C.

Multivariate Regression

Now suppose the outcome of interest is $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$.

- Use the notation $\mathrm{F}_j(\theta_j)$ for the distribution function of y_j, $j = 1, \ldots, p$.
- The joint distribution function can be expressed using a copula C as
$$F = C(F_1, \ldots, F_p).$$
- The set of parameters is
  - $\theta = \{\theta_1, \ldots, \theta_p, \alpha\}$,
  - where α is the set of parameters associated with the copula C.
- Copula functions work particularly well with continuous variables. There is less evidence about their utility for fitting discrete outcomes (or mixtures).
- It is customary, although not necessary, to let θ_j depend on explanatory variables $\mathbf{x}$ and to use constant α.

Student Achievement

- We now have a way to assess the joint distribution of the dependent variables by (1) specifying the marginal distributions and (2) the copula
- Consider a regression context of student assessment
- Data from the 1988 NELS. We consider a random sample of $n = 1,000$ students
  - Y_1= math score
  - Y_2= science score
  - Y_3= reading score
  - explanatory variables: minority, ses (socio-economic status), female, public, schoolsize, urban, and rural
- Some Summary Statistics

	meanSummary	medSummary	sdSummary	minSummary	maxSummary
read	26.6786	25.650	8.753	10.47	43.83
math	35.8115	34.280	12.091	17.27	66.59
sci	18.4148	17.920	4.955	9.56	32.88
ses	-0.0347	-0.037	0.814	-2.41	1.84
schoolsize	613.3000	500.000	339.158	100.00	1400.00

Student Achievement

- Student achievement scores are slightly right-skewed.
- Consider gamma regression

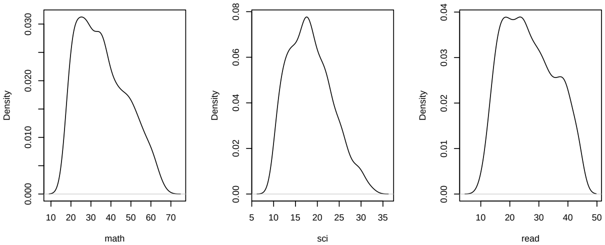


Figure : Student Achievement Scores. Somewhat skewed.

Comparing Achievement Scores

- Even after controlling for explanatory variables, scores are highly related.

```
> round(cor(cbind(umath,usci,uread),method = c("spearman")),digits=3)
      umath  usci uread
umath 1.000 0.651 0.646
usci  0.651 1.000 0.650
uread 0.646 0.650 1.000
```

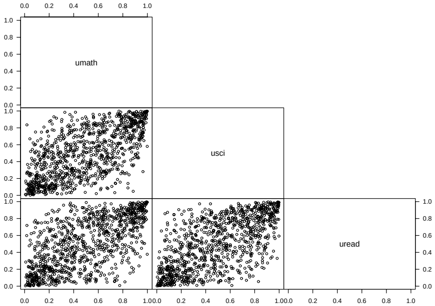


Figure : Scatterplot matrix of Prob Int Transformed student math, science and reading scores.

Likelihood Analysis

- We wish to estimate the full likelihood simultaneously
  - Using the chain-rule from calculus, we have

$$\begin{aligned}\frac{\partial^2}{\partial y_1 \partial y_2} \mathrm{F}(y_1, y_2) &= \frac{\partial^2}{\partial y_1 \partial y_2} \mathrm{C}(\mathrm{F}_1(y_1), \mathrm{F}_2(y_2)) \\ &= \mathrm{f}_1(y_1)\mathrm{f}_2(y_2)\mathrm{c}(\mathrm{F}_1(y_1), \mathrm{F}_2(y_2)),\end{aligned}$$

    where f_j and c are densities corresponding to the distribution functions F_j and C.
  - Taking logs, we have

$$L = \ln \mathrm{f}_1(y_1) + \ln \mathrm{f}_2(y_2) + \ln \mathrm{c}(\mathrm{F}_1(y_1), \mathrm{F}_2(y_2))$$

- F_1 - math - set of beta's, F_2 - sci - set of beta's
- one parameter for the copula

Comparison of Independence to Copula Models

	Coefficient	t-statistic	Coefficient	t-statistic
Copula Parameter			5.265	22.125
Math Parameters				
Intercept	3.605	137.077	3.611	137.233
minority	-0.074	-3.228	-0.060	-2.635
ses	0.174	12.980	0.176	13.089
female	0.015	0.766	0.015	0.795
public	-0.049	-1.845	-0.056	-2.150
school1	0.001	0.304	-0.001	-0.192
urban	0.015	0.620	0.011	0.461
rural	0.017	0.654	0.031	1.247
Science Parameters				
Intercept	2.985	140.746	2.987	120.422
minority	-0.104	-5.576	-0.089	-4.212
ses	0.113	10.434	0.110	8.620
female	-0.024	-1.559	-0.025	-1.421
public	-0.041	-1.911	-0.039	-1.611
school1	0.000	-0.128	-0.001	-0.297
urban	-0.011	-0.557	-0.017	-0.744
rural	0.007	0.340	0.019	0.789

Multivariate Regression

Several outcomes of interest (multivariate), several variables available to explain/predict these outcome (regression)

Why multivariate regression?

- Sharing of information - as with SUR (seemingly unrelated regressions). This is an efficiency argument - most helpful for small data sets.

Multivariate Regression

Several outcomes of interest (multivariate), several variables available to explain/predict these outcome (regression)

Why multivariate regression?

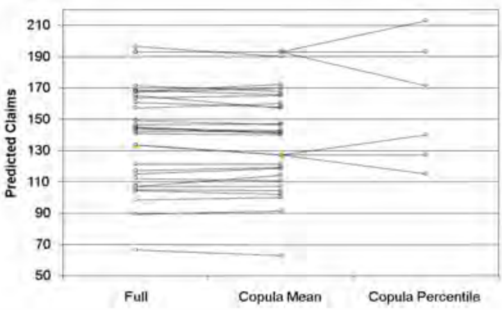
- Sharing of information - as with SUR (seemingly unrelated regressions). This is an efficiency argument - most helpful for small data sets.
- Scientific interest. The main purpose is to understand how outcomes are related. For example, when I control for claimant's age, gender, use of lawyer and so forth, how are losses and expenses related?
- Prediction. Assessing association is particularly important for the tails.
  - In the school example, the interest is in predicting the tails of the joint distribution. Which children are performing poorly (well) in math, science, and reading (simultaneously)?

Special Case: Longitudinal Data

- Here, we think of $\mathbf{y} = (y_1, \ldots, y_T)'$ as a short time series from an outcome of interest, e.g., commercial auto claims from a battery company.
- There is an extensive literature on linear longitudinal data models and their connections to credibility theory.
- More recently, many are working on generalized linear model outcomes with random effects (GLMMs) to handle extensions to medium/thick tail distributions.

Example. Massachusetts Auto Claims, NAAJ (2005), IME (2006)

- Consider claims arising from bodily injury liability.
- We have annual data from $n = 29$ towns over $T = 6$ years, 1993-1998, inclusive.
- On the margin, we used gamma regressions.
  - Two explanatory variables used for premium rating were (a) population per square mile (log units) and (b) per capita income
- A Gaussian copula was used for time dependencies



A More Complex Example. Singapore Auto Claims, JASA (2008), Astin (2008)

- Consider claims arising from three types of auto coverages.
  - bodily injury
  - own damage
  - third party claims
- Each is skewed and heavy-tailed

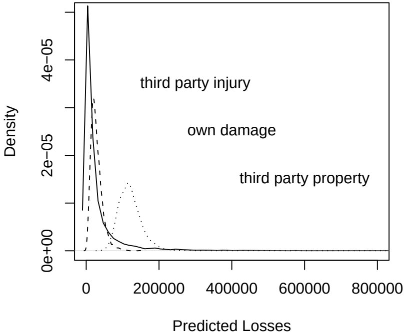


Figure : Density by Coverage Type

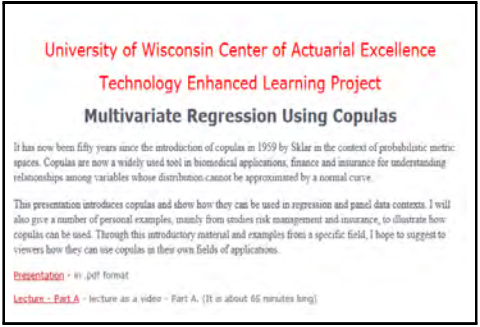
Singapore Auto Claims

- Data Features
  - Each policyholder may have 0, 1 or more (up to 5) claims.
  - Each claim yields one of 7 $(= 2^3 - 1)$ combinations of the three coverages
  - Lots of variables to explain the presence and extent of a claim (age, sex, driving history and so on)
- Model Features
  - Used a random effects Poisson for claim counts
  - A multinomial logit for claim type
  - A copula model with GB2 marginal regressions for claims severity
- Results - Important associations among coverage severities

Advertising 3

You can learn more about copula regression at our *Technology Enhanced Learning Project:*

`http://instruction.bus.wisc.edu/jfrees/UWCAELearn/default.aspx`



University of Wisconsin Center of Actuarial Excellence

Technology Enhanced Learning Project

Multivariate Regression Using Copulas

It has now been fifty years since the introduction of copulas in 1959 by Sklar in the context of probabilistic metric spaces. Copulas are now a widely used tool in biomedical applications, finance and insurance for understanding relationships among variables whose distribution cannot be approximated by a normal curve.

This presentation introduces copulas and show how they can be used in regression and panel data contexts. I will also give a number of personal examples, mainly from studies risk management and insurance, to illustrate how copulas can be used. Through this introductory material and examples from a specific field, I hope to suggest to viewers how they can use copulas in their own fields of applications.

Presentation - in .pdf format

Lecture - Part A - lecture as a video - Part A. (It is about 65 minutes long)

Why Multivariate Outcomes?

- For some products, insurers must track payments separately by component to meet contractual obligations.
  - In automobile coverage, deductibles and limits depend on the coverage type, e.g., bodily injury, damage to one's own vehicle or to another party.
  - In medical insurance, there are often co-pays for routine expenditures such as prescription drugs.
  - In personal lines umbrella insurance, there are separate limits for homeowners and auto coverages, as well overall limits for losses from all sources.

Why Multivariate Outcomes?

- For some products, insurers must track payments separately by component to meet contractual obligations.
  - In automobile coverage, deductibles and limits depend on the coverage type, e.g., bodily injury, damage to one's own vehicle or to another party.
  - In medical insurance, there are often co-pays for routine expenditures such as prescription drugs.
  - In personal lines umbrella insurance, there are separate limits for homeowners and auto coverages, as well overall limits for losses from all sources.
- For other products, there may be no contractual reasons to decompose but insurers do so anyway to better understand the risk, e.g., homeowners insurance.
- Multivariate models need not be restricted to only insurance losses, e.g., Example 3 study of term and whole life insurance ownership, or assets such as stocks and bonds.
- Commonly understood that $Uncertainty(Z_1+Z_2) \neq Uncertainty(Z_2)+Uncertainty(Z_2)$. Need to understand the joint behavior of risks (Z_1,Z_2).

Basic Notation of a Multivariate Two-Part Model

Use a multivariate outcome of interest $\mathbf{y}$ where each element of the vector consists of two parts. Thus, we observe

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \quad \text{as well as} \quad \mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_p \end{pmatrix}$$

and potentially observe

$$\mathbf{y}^* = \begin{pmatrix} y_1^* \\ \vdots \\ y_p^* \end{pmatrix}.$$

- $\mathbf{r}$ - the frequency vector, $\mathbf{y}^*$ as the amount, or severity, vector.
- Decompose the overall likelihood into frequency and severity components

$$\mathrm{f}(\mathbf{r},\mathbf{y}^*) = \mathrm{f}_1(\mathbf{r}) \times \mathrm{f}_2(\mathbf{y}^*|\mathbf{r})$$

Let's look at some **Actuarial Applications**.

Example 1. Health Care Expenditures, *NAAJ* (2011)

- Medical Expenditure Panel Survey
  - 9,472 participants from 2003 for in-sample, 9,657 participants from 2004 for validation
- p=2 Outcomes of Interest - Inpatient (Hospital) and Outpatient Expenditures
- Explanatory Variables - About 30. Includes demography (age, sex, ethnicity), socio-economic (education, marital status, income), health status, employment (status, industry), health insurance
- Frequency Model - Logistic, Negative Binomial models
- Severity Model - Gamma regression, mixed linear models

Example 2. Multi-Peril Homeowners Insurance

Table : Summarizing 404,664 Policy-Years, $p = 9$ Perils

Peril (j)	Frequency (in percent)	Number of Claims	Median Claims
Fire	0.310	1,254	4,152
Lightning	0.527	2,134	899
Wind	1.226	4,960	1,315
Hail	0.491	1,985	4,484
WaterWeather	0.776	3,142	1,481
WaterNonWeather	1.332	5,391	2,167
Liability	0.187	757	1,000
Other	0.464	1,877	875
Theft-Vandalism	0.812	3,287	1,119
Total	5.889*	23,834*	1,661

Example 2. Multi-Peril Homeowners Insurance

- Work appeared in *Astin Bulletin* (2010) and *Variance* 2013
- We drew two random samples from a homeowners database maintained by the Insurance Services Office.
  - This database contains over 4.2 million policyholder years.
  - Policies issued by several major insurance companies in the United States, thought to be representative of most geographic areas in the US.
- Our in-sample, or "training," dataset consists of a representative sample of 404,664 records taken from this database.
  - We estimated several competing models from this dataset
- We use a held-out, or "validation" subsample of 359,454 records, whose claims we wish to predict.

Multi-Peril Homeowners Insurance Results

- Model
  - Outcomes of interest (p=9)
  - Explanatory variables - over 100. Proprietary information collected by Insurance Services Office.
  - Frequency Model - Dependence ratio models of multivariate binary data
  - Severity Model - Gaussian copula with gamma regression marginals
- Results
  - We established strong dependencies in the frequencies. The dependence ratio model (Ekholm et al, Biometrika, 1995) was helpful.
  - For severities, insufficient number of joint claims within a year to see strong dependencies.

Example 3. Life Insurance Ownership, *NAAJ* 2010

- Survey of Consumer Finances - 2,150 households from the 2004 survey
- p=2 Outcomes of interest, amount of term life insurance and the net amount at risk for whole life insurance
  - Substitute or Complement?
- Explanatory Variables - About 30. Includes assets, debt, income, bequests and inheritance, age, education and financial vulnerability
- Frequency Model - Bivariate probit
- Severity Model - Gaussian copula with gamma regression marginals

Example 4. Healthcare Expenditures

- Work to appear in the *Annals of Actuarial Science* (2013)
- From the 2006 Medical Expenditure Panel Survey (MEPS), $n = 18,908$ individuals
- There are $p = 5$ expenditure types

Summary Statistics of Expenditures by Event Types

Event Type	Count*	Percent	Mean	Standard Deviation	Median	Maximum
Office-Based (OB)	10,528	55.7	1,653	5,336	420	199,696
Hospital Outpatient (OP)	2,164	11.4	2,817	7,517	909	256,741
Emergency Room (ER)	2,274	12.0	1,311	2,398	566	33,412
Inpatient (IP)	1,339	7.1	16,604	36,133	7,548	693,483
Home Health (HH)	235	1.2	14,092	36,611	3,312	394,913

* An observation is a person who has this type of medical event during the year.

- Not surprisingly, different event types have very different expenditure distributions

Counts by Type of Event

Singles		Doubles		Triples		Quadruples	
OB	6703	OB,OP	1303	OB,OP,ER	279	OB,OP,ER,IP	163
OP	130	OB,ER	917	OB,OP,IP	154	OB,OP,ER,HH	11
ER	396	OB,IP	427	OB,OP,HH	26	OB,OP,IP,HH	27
IP	59	OB,HH	63	OB,ER,IP	362	OB,ER,IP,HH	28
HH	7	OP,ER	30	OB,ER,HH	14	OP,ER,IP,HH	0
		OP,IP	10	OB,IP,HH	32		
		OP,HH	2	OP,ER,IP	9		
		ER,IP	44	OP,ER,HH	0		
		ER,HH	1	OP,IP,HH	1		
		IP,HH	3	ER,IP,HH	1		
Subtotal	7,295		2,800		878		229

* There are 7,687 observations without any events during the year.
* There are 19 observations with all five events during the year.

- There are many joint event occurrences - more so than suggested by a model of independence
- Many more triples, quadruples, quintuples than suggested by a model of independence
- We consider higher order dependencies via *conditional* odds ratios.

- Multinomial Logistic Regressions
  - With the binary vector $\mathbf{r}$, there are 2^p possible events
- Marginal Binary Regressions
  - With a marginal logistic regression model, we would employ

$$\Pr(r_{i1}=1)=\pi_{i1}=\frac{\exp(\mathbf{x}_{i1}'\beta_1)}{1+\exp(\mathbf{x}_{i1}'\beta_1)},$$

    resulting in $\mathrm{logit}(\pi_{i1})=\mathbf{x}_{i1}'\beta_1$.
  - Explanatory variables ($\mathbf{x}$) and regression coefficients (β) to depend on the type of outcome.

Fit Five Marginal Logistic Regression Models

Table 13: Marginal Logistic Regressions for Five Types of Events

Category	Variable	Office-Based			Hospital Outpatient			Emergency Room			Inpatient			Home Health		
		Estimate	t-value		Estimate	t value		Estimate	t value		Estimate	t value		Estimate	t value	
	(Intercept)	-2.463	-23.798	***	-5.219	-31.313	***	-2.224	-16.939	***	-4.279	-23.343	***	-8.968	-17.865	***
Demography	AGE	0.007	4.046	***	0.015	6.352	***	-0.019	-9.295	***	-0.007	-2.895	**	0.030	4.415	***
	GENDER	0.584	16.167	***	0.398	7.935	***	0.201	4.188	***	0.563	8.629	***	0.246	1.672	.
Ethnicity	ASIAN	-0.492	-5.853	***	-0.486	-3.163	**	-0.621	-3.764	***						
	BLACK	-0.438	-9.118	***	-0.260	-3.717	***	0.155	2.625	**						
Region	NORTHEAST	0.041	0.735		0.598	7.706	***	0.214	2.748	**	0.117	1.186				
	MIDWEST	0.139	2.733	**	0.585	8.071	***	0.310	4.453	***	0.249	2.825	**			
	SOUTH	0.044	0.996		0.197	2.855	**	0.196	3.142	**	0.197	2.578	**			
Access to Care	USC	1.292	32.451	***	0.822	10.352	***	0.362	6.020	***	0.422	5.189	***			
Education	HIGHSCH	0.104	2.310	*	0.235	3.590	***				0.121	1.628				
	COLLEGE	0.289	5.338	***	0.219	2.979	**				0.302	3.310	***			
Marital Status	MARRIED	0.248	5.093	***	0.209	2.762	**							-1.022	-5.314	***
	WIDIVSEP	0.144	2.337	*	0.128	1.505		0.258	4.175	***				-0.503	-2.675	**
Family Size	FAMSIZE	-0.114	-9.749	***	-0.076	-4.090	***	-0.038	-2.519	*	0.058	3.201	**			
Income	HINCOME	0.131	2.059	*				-0.378	-5.148	***	-0.586	-5.919	***	-0.229	-1.008	
	MINCOME	0.063	1.084					-0.445	-6.404	***	-0.417	-4.664	***	-0.518	-2.464	*
	LINCOME	-0.072	-1.143					-0.241	-3.183	**	-0.338	-3.486	***	-0.372	-1.730	.
	NPOOR	-0.115	-1.363					-0.093	-0.927		-0.080	-0.650		-0.284	-1.063	
Physical Health	POOR	1.308	10.535	***	1.108	9.789	***	1.144	10.122	***	1.184	9.044	***	1.449	3.914	***
	FAIR	1.116	15.785	***	0.863	9.638	***	0.850	9.826	***	0.706	6.438	***	1.294	3.630	***
	GOOD	0.587	12.507	***	0.488	6.421	***	0.425	5.971	***	0.409	4.355	***	0.550	1.522	
	VGOOD	0.421	9.413	***	0.262	3.443	***	0.224	3.123	**	0.228	2.407	*	0.663	1.799	.
Mental Health	MNHPOOR	0.301	3.792	***				0.125	1.608							
Any Limitation	ANYLIMIT	0.777	15.741	***	0.726	12.652	***	0.574	9.918	***	0.631	8.617	***	2.051	9.032	***
Unemployment	UNEMPLOYED	0.139	3.130	**							0.482	6.441	***	1.244	5.481	***
Industry	EDUCHEALTH	0.105	2.021	*							0.160	1.702	.	0.745	2.602	**
	PUBADMIN	0.208	2.251	*							0.336	2.064	*			
	NATRESOURCE							-0.723	-2.300	*	-0.829	-1.621				
Insurance	INSURED	0.672	13.202	***	0.662	7.949	***	0.297	5.027	***	0.532	6.673	***	1.424	5.364	***
Managed Care	MANAGEDCARE	0.157	3.540	***	0.138	2.371	*									
Model fit indices	AIC		20,734.74			11,785.83			12,989.63			8,834.53			1,851.78	
	Log-Likelihood		-10,338.37			-5,871.92			-6,471.82			-4,393.27			-908.89	

Dependence Ratios

- Correlations are linear measures – they do not capture associations well for binary data
- One association measure for binary variables is the *dependence ratio*,

$$\tau_{12} = \frac{\Pr(r_1 = 1, r_2 = 1)}{\Pr(r_1 = 1)\Pr(r_2 = 1)},$$

  the ratio of the joint probability to the product of the marginal probabilities.
  - In the case of independence, τ_{12} to be 1.
  - Values of $\tau_{12} > 1$ indicate positive dependence
  - Values of $\tau_{12} < 1$ indicate negative dependence.

Joint Counts Among Event Types

	OB	OP	ER	IP	HH
Office-Based (OB)	-	1,982	1,793	1,212	220
Hospital Outpatient (OP)	1,982	-	511	383	86
Emergency Room (ER)	1,793	511	-	626	74
Inpatient (IP)	1,212	383	626	-	111
Home Health (HH)	220	86	74	111	-
Total Count for the Event	10,528	2,164	2,274	1,339	235
Percent of an Event	55.7	11.4	12.0	7.1	1.2
Odds of an Event	1.256	0.129	0.137	0.076	0.013

- Office-Based (OB) expenditures are the most prevalent
- Home Health (HH) expenditures are the least prevalent

Dependence Ratios Among Event Types

	OB	OP	ER	IP	HH
Office-Based (OB)	-				
Hospital Outpatient (OP)	1.645	-			
Emergency Room (ER)	1.416	1.963	-		
Inpatient (IP)	1.626	2.499	3.887	-	
Home Health (HH)	1.681	3.198	2.618	6.670	-

- The large dependency ratio between HH and IP (6.67) indicates a large positive association
- The small dependency ratio between ER and OB (1.416) indicates a small positive association

Odds Ratios

- Another approach – using odds ratios
- recall that the *odds* of $\{r_1 = 1\}$ is

$$odds(r_1) = \frac{\pi_1}{1-\pi_1} = \frac{\Pr(r_1=1)}{\Pr(r_1=0)}.$$

- The odds ratio between r_2 and r_1 is

$$\begin{aligned} \mathrm{OR}(r_2, r_1) &= \frac{odds(r_2|r_1=1)}{odds(r_2|r_1=0)} \\ &= \frac{\Pr(r_2=1, r_1=1)\Pr(r_2=0, r_1=0)}{\Pr(r_2=1, r_1=0)\Pr(r_2=0, r_1=1)}. \end{aligned}$$

- The odds ratio is one under independence. Values greater than one indicate positive dependence and values less than one indicate negative dependence.

Odds Ratios Among Types of Events

	OB	OP	ER	IP	HH
Office-Based (OB)	-				
Hospital Outpatient (OP)	10.447	-			
Emergency Room (ER)	3.371	2.627	-		
Inpatient (IP)	8.454	3.551	8.482	-	
Home Health (HH)	11.902	4.609	3.442	12.717	-

- The large odds ratio between HH and IP (12.717) indicates a large positive association
- The small odds ratio between ER and OB (3.371) indicates a small positive association
  - This is consistent with the dependency ratio approach
  - Smallest odds ratio is between OP and ER, smallest dependence ratio is between ER and OB – they measure different aspects of association.

Marginal Regression Results

- Fits of marginal regressions confirm that covariates have an important influence on the type and amount of expenditure
- Marginal frequency fits are based on logistic regressions with selected covariates for each event type
- However, these marginal regression fits to do not account for the association among types

Table : Association Test Statistics From Logistic Regression Fits

	OB	OP	ER	IP	HH
Office-Based (OB)	-				
Hospital Outpatient (OP)	10.735	-			
Emergency Room (ER)	9.124	8.444	-		
Inpatient (IP)	10.313	9.979	25.471	-	
Home Health (HH)	1.943	2.758	1.763	10.072	-

Odds Ratio Estimates

- We used the odds ratio approach is used to model the multivariate binary frequencies.
- Estimation is based on a likelihood approach where the likelihood is written in terms of marginal probabilities and odds ratios (e.g., Liang, Qaqish and Zeger, 1992).

		Empirical Estimate without Covariates	Likelihood Estimates with Covariates	t-statistic
Bivariate	OB, OP	10.447	5.603	9.790
	OB, ER	3.371	2.905	10.741
	OB, IP	8.454	6.669	8.378
	OP, ER	2.627	1.985	8.059
	OP, IP	3.551	2.532	8.610
	ER, IP	8.482	6.444	13.186
Triple	OB, OP, ER	0.357	0.438	-6.106
	OB, OP, IP	0.218	0.324	-7.443
	OB, ER, IP	0.437	0.397	-7.149
	OP, ER, IP	0.500	0.560	-5.297
Quadruple	OB, OP, ER, IP	2.916	2.075	0.851

Odds Ratio Estimates

- Estimates of associations are, not surprisingly, strongly statistically significant, even after controlling for covariates
- We also used a copula regression model for expenditure amounts. Dependencies were not as strong as with frequencies.
- But does this model mean that we have better predictions of medical expenditures??

Advertising 4

- You will be able to learn more about predictive modeling at our *Predictive Modeling Applications in Actuarial Science* project at the website `http://instruction2.bus.wisc.edu/course/view.php?id=8`
- This is a two-volume series that will be published by *Cambridge University Press*.
- Co-Editors: EW (Jed) Frees, Richard Derrig, Glenn Meyers
- Many important authors here, including Katrien Antonio, Greg Taylor, Peng Shi, Montseratt Guillen, others(?)
- Volume 1 is foundations, Volume 2 is case studies
- Co-sponsored by the Casualty Actuarial Society and the Canadian Institute of Actuaries

Predictive Modeling Applications in Actuarial Science

1. Introduction to Predictive Modeling in Actuarial Science	Jed Frees, Glenn Meyers , Richard Derrig
2. Multiple Linear Regression	Margie Rosenberg, Jim Guszcza
3. Regression with Categorical Dependent Variables	Montserratt Guillen
4. Regression with Count Dependent Variables	Jean-Philippe Boucher
5. Generalized Linear Models	Gary Dean
6. Frequency/Severity Models	Winnie Sun
7. Mixed Models	Katrien Antonio, Wayne Zhang
8. Generalized Additive Models, including Non-Parametric Regression	Pat Brockett
9. Fat-Tail Regression Models	Peng Shi
10. Spatial Statistics	Claudia Czado, Eike Brechmann
11. Unsupervised Learning.	Louise Frances
12. Bootstrapping & Simulation	Ken Seng Tan
13. Introduction to Bayesian Computational Methods	Brian Hartman
14. Bayesian Regression Models	Enrique de Alba, Luis E. Nieto-Barajas
15. Time Series, including Lee-Carter forecasting	Piet de Jong
16. Longitudinal and Panel Data Models	Jed Frees
17. Credibility and Regression Modeling	Vytaras Brazauskas, Harald Dornheim, Ponmalar Ratnam
18. Survival Models, including Cox Regression	Jim Robinson
19. Claims Triangles/Loss Reserves	Greg Taylor
20. Transition Modeling	Bruce Jones

Table of Contents

Predictive Models

Here are some skills/topics useful in predictive modeling

- Two-Part? For example, loss or no loss
- Loss distributions are typically skewed and heavy-tailed
- Censored?
  - Losses censored by amounts through deductibles or policy limits
  - Loss censored by time, e.g., claim triangles
- Insurance data typically has lots of explanatory variables. Lots.
- Multivariate responses, e.g., types of coverages, perils, bundling of insurances
- Longitudinal (panel)? Are you following the contract over time?
- Losses credible? we often wish to incorporate external knowledge into our analysis

Predictive Models

Here are some skills/topics useful in predictive modeling

- Two-Part? For example, loss or no loss
- Loss distributions are typically skewed and heavy-tailed
- Censored?
  - Losses censored by amounts through deductibles or policy limits
  - Loss censored by time, e.g., claim triangles
- Insurance data typically has lots of explanatory variables. Lots.
- Multivariate responses, e.g., types of coverages, perils, bundling of insurances
- Longitudinal (panel)? Are you following the contract over time?
- Losses credible? we often wish to incorporate external knowledge into our analysis
- Spatial? e.g., hurricane, flood data
- High Frequency Observations?, e.g., telematics, usage data

Gini - Research Motivation

- We are proposing several new methods of determining premiums (e.g., instrumental variables, copula regression)
  - How to compare?
  - No single statistical model that could be used as an “umbrella” for likelihood comparisons

Gini - Research Motivation

- We are proposing several new methods of determining premiums (e.g., instrumental variables, copula regression)
  - How to compare?
  - No single statistical model that could be used as an “umbrella” for likelihood comparisons
- Want a measure that not only looks at statistical significance but also monetary impact
- Would like a measure to help distinguish among premiums – when is it advantageous for an insurer to introduce a refined rating system?

Premiums and Relativities

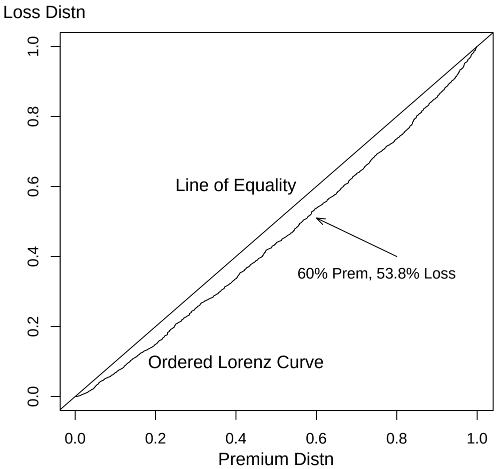
- The relativity is $R(\mathbf{x}_i) = S(\mathbf{x}_i)/P(\mathbf{x}_i)$.
- Suppose that the score S is a good approximation to the expected loss
  - If the relativity is small, then we can expect a small loss relative to the premium. This is a profitable policy but is also one that is susceptible to potential raiding by a competitor (adverse selection)
  - If the relativity is large, then we can expect a large loss relative to the premium. This is one where better loss control measures, e.g., renewal underwriting restrictions, can be helpful.

Premiums and Relativities

- The relativity is $R(\mathbf{x}_i) = S(\mathbf{x}_i)/P(\mathbf{x}_i)$.
- Suppose that the score S is a good approximation to the expected loss
  - If the relativity is small, then we can expect a small loss relative to the premium. This is a profitable policy but is also one that is susceptible to potential raiding by a competitor (adverse selection)
  - If the relativity is large, then we can expect a large loss relative to the premium. This is one where better loss control measures, e.g., renewal underwriting restrictions, can be helpful.
- Through the relativities, we can form portfolios of policies and compare losses to premiums to assess profitability
- This is the goal of the *ordered* Lorenz curve that we introduce

Homeowners Example

- Here is an ordered Lorenz curve
- To summarize the separation, the Gini index is 10.03% with a standard error of 1.45% .



Ordered Lorenz Curve

- Notation
  - $\mathbf{x}_i$ - explanatory variables, $P(\mathbf{x}_i)$ - premium, y_i - loss, $R_i = R(\mathbf{x}_i)$, $\mathrm{I}(\cdot)$ - indicator function, and $\mathrm{E}(\cdot)$ - mathematical expectation
- The Ordered Lorenz Curve
  - Vertical axis

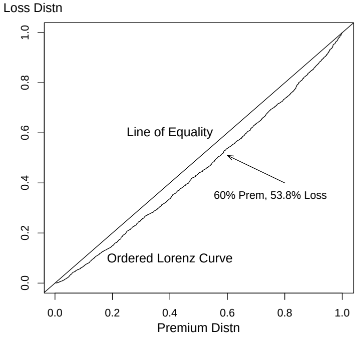
$$F_L(s) = \frac{\mathrm{E}[y\mathrm{I}(R \leq s)]}{\mathrm{E}\ y} \underset{empirical}{=} \frac{\sum_{i=1}^n y_i \mathrm{I}(R_i \leq s)}{\sum_{i=1}^n y_i}$$

    the proportion of losses.
  - Horizontal axis

$$F_P(s) = \frac{\mathrm{E}[P(\mathbf{x})\mathrm{I}(R \leq s)]}{\mathrm{E}\ P(\mathbf{x})} \underset{empirical}{=} \frac{\sum_{i=1}^n P(\mathbf{x}_i) \mathrm{I}(R_i \leq s)}{\sum_{i=1}^n P(\mathbf{x}_i)}$$

    the proportion of premiums.

- To read the ordered Lorenz Curve
  - Pick a point on the horizontal axis, say 60% of premiums
  - The corresponding vertical axis is about 53.8% of losses
  - This represents a profitable situation for the insurer
  - The “line of equality” represents a break-even situation
- To summarize the separation, the Gini index is 10.03% with a standard error of 1.45% .



Thinking About our Gini Index

- Definition - The Gini as an area

$$Gini = 2 \int_0^\infty \{F_P(s) - F_L(s)\}\, dF_P(s).$$

- From this, interpret the Gini index as a measure of profit

$$\frac{1}{n} \sum_{i=1}^{n} \left(\hat{F}_P(R_i) - \hat{F}_L(R_i) \right) \approx \frac{\widehat{Gini}}{2},$$

  - It is an “average profit” in the sense that we are taking a mean over all decision-making strategies, that is, each strategy retaining the policies with relativities less than or equal to R_i.
  - Insurers that adopt a rating structure with a large Gini index are more likely to enjoy a profitable portfolio.

Gini Interpretation

- We show that

$$Gini \approx \frac{2}{n} \widehat{\mathrm{Cov}}\left((y-P), rank(R)\right).$$

- Think about $P-y$ as the “profit” associated with a policy.
- The Gini index is proportional to the negative covariance between profits and the rank of relativities.
  - If policies with low profits$\sim$high relativities and high profits$\sim$low relativities, then the Gini index is positive and large.

Gini Interpretation

- We show that

$$Gini \approx \frac{2}{n} \widehat{\mathrm{Cov}}\left((y - P), rank(R) \right).$$

- Think about $P - y$ as the “profit” associated with a policy.
- The Gini index is proportional to the negative covariance between profits and the rank of relativities.
  - If policies with low profits$\sim$high relativities and high profits$\sim$low relativities, then the Gini index is positive and large.
- **Predictive Modeling Application:** The Gini index gives us another summary statistic for measuring the relationship between premium estimates calibrated on in-sample data with out-of-sample losses.

Gini Summary

- The ordered Lorenz curve allows us to capture the separation between losses and premiums in an order that is most relevant to potential vulnerabilities of an insurer's portfolio
  - The corresponding Gini index captures this potential vulnerability
- We have introduced measures to quantify the statistical significance of empirical Gini coefficients
  - The theory allows us to compare different Ginis
  - It is also useful in determining sample sizes

MEPS Model Validation

- The paper uses 2006 healthcare expenditures for in-sample data
- The paper uses 2007 expenditures for out-sample data (different people)
- Compared many models on an out-of-sample basis
  - Examined frequency/severity and pure premiums
  - Looked at univariate and multivariate (over expenditure types)
  - Also analyzed multinomial logits for frequencies

MEPS Model Validation

- The paper uses 2006 healthcare expenditures for in-sample data
- The paper uses 2007 expenditures for out-sample data (different people)
- Compared many models on an out-of-sample basis
  - Examined frequency/severity and pure premiums
  - Looked at univariate and multivariate (over expenditure types)
  - Also analyzed multinomial logits for frequencies
- Table compares several scoring measures (rows), several decision criteria (columns)
  - Difficult to choose among alternative models

Table 15: Out-of-Sample Statistics

	Mean Absolute Error	Mean Absolute Percentage Error	Root Mean Square Error	Correlations Pearson	Spearman	Gini
One Part Models						
BasicOnePart	3,874.938	227.540	13,774.179	25.680	44.688	18.296
LogOnePart	2,684.980	6,993.282	14,300.376	22.565	51.515	18.608
SmearOnePart	10,934.435	174.691	35,244.110	22.565	51.515	18.608
Tweedie	3,589.482	150.285	13,734.769	26.646	49.616	18.896
Two Part Models						
BasicTPM	3,665.951	329.365	13,740.511	26.621	45.791	18.498
TPMLogNSev	2,705.836	525.740	14,209.043	27.036	50.521	18.896
TPMSmearSev	3,630.774	160.093	13,718.385	27.036	50.521	18.896
TPMGammaSev	3,579.156	156.046	13,720.311	27.091	50.109	18.893
Multivariate One Part Models						
INDBasicOnePart	3,874.938	227.540	13,774.179	25.680	44.688	18.296
INDLogOnePart	2,719.334	9,387.637	14,433.423	22.584	51.474	18.350
INDOnePartTweed	2,781.582	9,787.627	14,513.104	21.256	47.050	18.376
INDBasicOnePartReduced	3,863.703	211.227	13,773.980	25.694	45.006	18.399
INDLogOnePartReduced	2,719.343	9,415.450	14,433.430	22.583	51.474	18.353
INDOnePartTweedReduced	2,781.683	9,526.059	14,513.085	22.570	46.912	18.509
Multivariate Two Part Models						
INDBasicTPM	3,603.476	265.246	13,731.740	26.811	48.913	18.940
INDTPMLogNSev	2,851.499	258.909	14,075.163	26.351	49.191	18.710
INDTPMGammaSev	3,579.410	146.869	13,736.173	26.677	49.794	18.974
INDBasicTPMReduced	3,597.650	192.595	13,732.308	26.803	48.663	18.834
INDTPMLogNSevReduced	2,848.996	258.562	14,062.045	26.879	49.856	18.743
INDTPMGammaSevReduced	3,574.799	149.014	13,726.846	26.883	50.184	18.956
CellTPMLogNSev	3,053.903	185.672	14,013.866	25.429	50.036	18.773
DepTPMLogNSevReduced	2,822.43	272.728	14,082.72	26.968	50.058	18.856
DepTPMGammaSevReduced	3,521.059	152.423	13,732.81	26.784	50.232	18.997

Predictive Distributions

- These measures only look at the forecast at a certain point, not the *distribution* of forecasts
- Using simulation, it is straight forward in principle to calculate the predictive distribution of the 2007 portfolio for any of these models (Important Bayesian principle)

Predictive Distributions

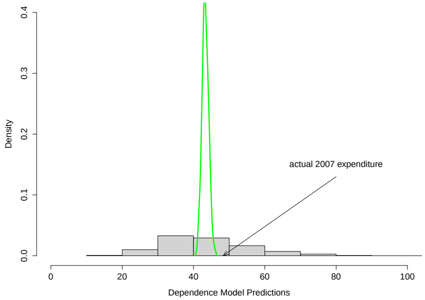
- These measures only look at the forecast at a certain point, not the *distribution* of forecasts
- Using simulation, it is straight forward in principle to calculate the predictive distribution of the 2007 portfolio for any of these models (Important Bayesian principle)
- Algorithm: For each replication in a large number of simulations (we used 1,000), we:
  - Use the predictor variables and the regression coefficient estimates to create parameter estimates of the distribution for each person in the 2007 portfolio
  - Simulate the expenditure from this distribution for each person in the 2007 portfolio
  - Summed expenditures over all persons in the 2007 portfolio
- This allows us to calculate the simulated distribution of losses

Predictive Distributions

- These measures only look at the forecast at a certain point, not the *distribution* of forecasts
- Using simulation, it is straight forward in principle to calculate the predictive distribution of the 2007 portfolio for any of these models (Important Bayesian principle)
- Algorithm: For each replication in a large number of simulations (we used 1,000), we:
  - Use the predictor variables and the regression coefficient estimates to create parameter estimates of the distribution for each person in the 2007 portfolio
  - Simulate the expenditure from this distribution for each person in the 2007 portfolio
  - Summed expenditures over all persons in the 2007 portfolio
- This allows us to calculate the simulated distribution of losses
- We compared the multivariate frequency-severity model assuming independence to our dependence model

MEPS Model Validation

- The actual 2007 value for the portfolio was $48.34 millions
- Green is the predictive distribution from the independence model
- Gray is the predictive distribution from the dependence model
  - Unlikely to occur in the model of independence although very plausible in the dependence model.
  - Does not validate the model of dependence but it is consistent with what we learned from our detailed in-sample analysis



Out-of-Sample Risk Measures

- The model of dependence exhibited a much wider predictive distribution of the held-out sample than the models that assume independence
- You may be more familiar with the following risk measures to establish this result
  - The $CTE(\alpha)$ is the expected value conditional on exceeding the $VaR(\alpha)$.
  - The VaR is simply a quantile or percentile, the $VaR(\alpha)$ gives the 100(1 - α) percentile of the distribution.

Risk Measure	Percentile 0.50	0.75	0.90	0.95	0.98	0.99	1.00
VaRInd	45.05	45.63	46.18	46.56	46.87	47.32	47.44
VaRDep	41.95	50.51	58.87	65.06	68.99	73.22	79.43
CTEInd	45.76	46.20	46.67	46.97	47.25	47.55	47.70
CTEDep	52.40	58.89	66.41	71.31	75.82	81.71	87.11

Predictive Modeling and Healthcare

- The goals and techniques of healthcare applications of predictive analytics are very similar to those of insurance company operations.
- Predictive analytics are used by
  - healthcare agencies,
  - managed care companies,
  - physicians, and others
- One main difference. In healthcare, there is a much greater emphasis on "loss control" and risk management
  - Using predictive models, managers will be able to target the most actionable patients who will benefit from targeted outreach and education.

Concluding Remarks

- Many short-term insurance losses come in two parts. It is common practice to examine frequency-severity modeling using regression covariates.
- Our contribution is focused on dependencies among losses. In recent years, this has become widely recognized as important in actuarial practice at many levels.
- For continuous severity modeling, copula regression appears the most promising. For discrete frequency distributions such as our binary outcomes, there are many more choices.

Concluding Remarks

- Research Theme: how to use micro-level data to make sensible statements about “macro-effects.”
- Certainly not the first to support this viewpoint
  - Traditional actuarial approach is to development life insurance company policy reserves on a policy-by-policy basis.
  - See, for example, Richard Derrig and Herbert I Weisberg (1993) “Pricing auto no-fault and bodily injury coverages using micro-data and statistical models”
- However, the idea of using voluminous data that the insurance industry captures for making managerial decisions is becoming more prominent.
  - Gourieroux and Jasiak (2007) have dubbed this emerging field the “microeconometrics of individual risk.”

Concluding Remarks

- In rating and reserving, actuaries have focused on “macro” approaches
  - Develop methods that have withstood the test of time
  - Compared to micro approaches, they are less expensive to implement
  - Compared to micro approaches, they are easier to interpret and explain
  - In contrast to Wall Street analysts, insurance analysts have not let the model drive their industry and used business knowledge to drive decision-making

Concluding Remarks

- In rating and reserving, actuaries have focused on “macro” approaches
  - Develop methods that have withstood the test of time
  - Compared to micro approaches, they are less expensive to implement
  - Compared to micro approaches, they are easier to interpret and explain
  - In contrast to Wall Street analysts, insurance analysts have not let the model drive their industry and used business knowledge to drive decision-making
- There are important strengths of a micro approach that focuses on the individual policyholder or claim
  - Easy to consider the effects of a changing mix of business
  - Readily allows for economic modeling of individual decision-making, e.g., what are the effects of claims reporting on changes in deductibles?
  - Allows us to considered “bundled” risks in a multivariate framework

Conclusion

Overheads are available at:
`https://sites.google.com/a/wisc.edu/jed-frees/`

Thank you for your kind attention.