Insurance Analytics

2019 Insurance Risk Analytics Symposium

Edward W. (Jed) Frees

University of Wisconsin – Madison

May 2019

Outline

Relevance of Insurance

By almost any measure, insurance is a major economy activity

- On a global level, insurance premiums comprised about 6.3% of the world gross domestic product (GDP) in 2013 (*Source*: International Insurance Fact Book, 2015)
  - Premiums accounted for 11.2% of GDP in Japan
  - Represented 7.5% of GDP in the United States
- On a personal level:
  - Almost everyone owning a home has insurance to protect themselves in the event of a fire, hailstorm, or some other calamitous event
  - Almost every country requires insurance for those driving a car

What is Analytics?

To set the stage, some ideas from my 2015 review paper in the *Annual Review of Financial Economics* entitled *Analytics of Insurance Markets*

- Insurance is a data-driven industry – analytics is a key to deriving information from data.
- But what is analytics?

What is Analytics?

To set the stage, some ideas from my 2015 review paper in the *Annual Review of Financial Economics* entitled *Analytics of Insurance Markets*

- Insurance is a data-driven industry – analytics is a key to deriving information from data.
- But what is analytics? Some alternative descriptors:
  - business intelligence may focus on processes of collecting data, often through databases and data warehouses
  - business analytics utilizes tools and methods for statistical analyses of data
  - data science can encompass broader applications in many scientific domains

What is Analytics?

To set the stage, some ideas from my 2015 review paper in the *Annual Review of Financial Economics* entitled *Analytics of Insurance Markets*

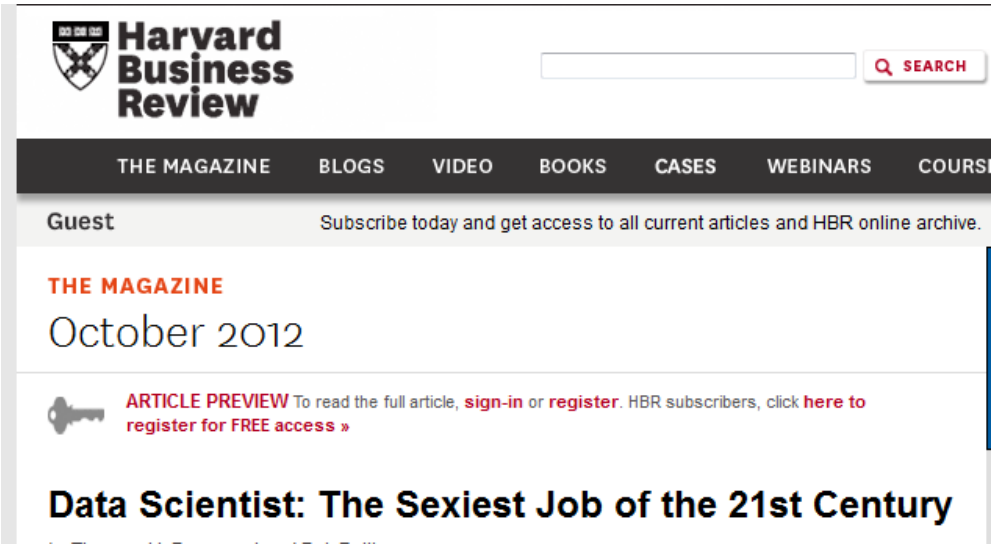
- Insurance is a data-driven industry – analytics is a key to deriving information from data.
- But what is analytics? Some alternative descriptors:
  - business intelligence may focus on processes of collecting data, often through databases and data warehouses
  - business analytics utilizes tools and methods for statistical analyses of data
  - data science can encompass broader applications in many scientific domains
- Analytics – the process of using data to make decisions.
  - This process involves gathering data, understanding models of uncertainty, making general inferences, and communicating results.

What is Analytics?

- Led by statistician W. Edwards Deming, an earlier generation sought to utilize quality improvement techniques to improve business processes, resulting in the field now known as *total quality management*.

What is Analytics?

- Led by statistician W. Edwards Deming, an earlier generation sought to utilize quality improvement techniques to improve business processes, resulting in the field now known as *total quality management*.
- Analytics continues to enjoy increasing popularity among businesses.

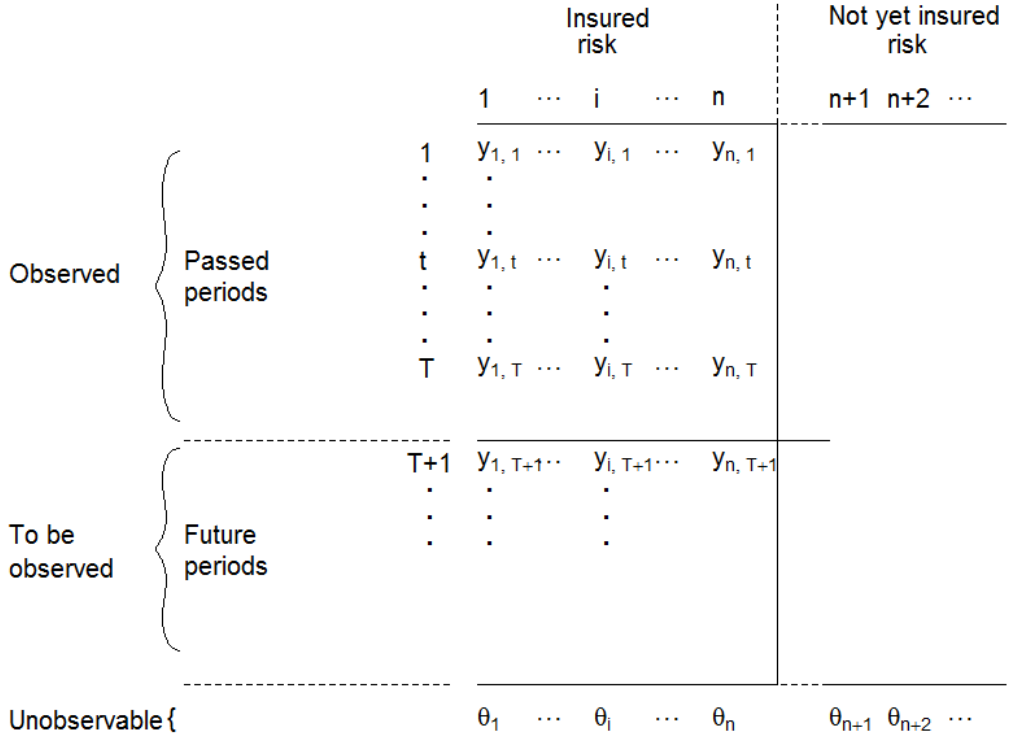


Why Predictive?

Why Predictive?

- Statisticians think about the traditional triad of inference: hypothesis testing, parameter estimation, and prediction.
- In insurance, predictions are useful for existing risks in future periods as well as not yet observed risks in a current period

Figure: Predictive Features of Insurance Analytics, Norberg (1979).



Impact of Analytics on our Teaching

- Historically, we focused on *in-sample fits*
  - Looked to summary measures such as R^2, AIC, BIC
- Now, the emphasis is on *predictive* ability of a model
  - For example, focus on *cross-validation* measures

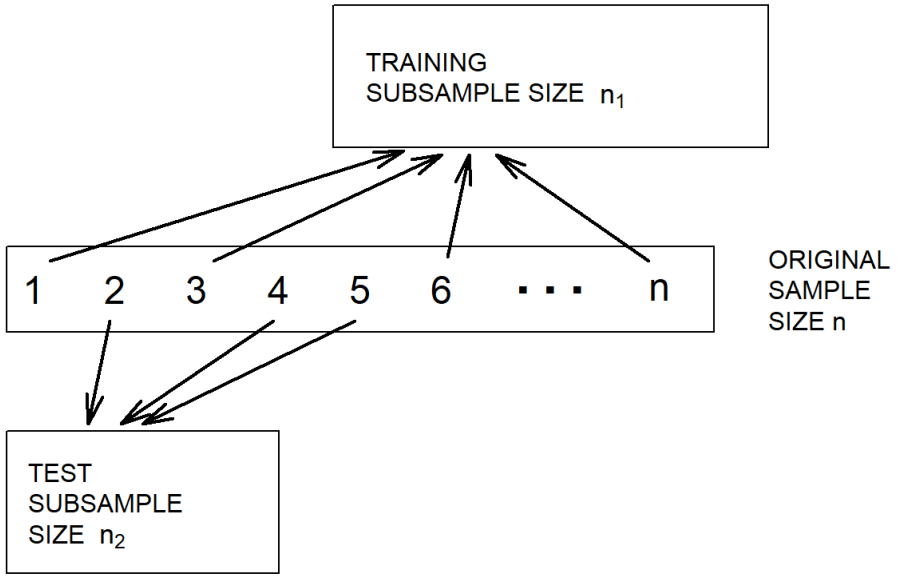


Figure: Model Validation. A data set is randomly split into two subsamples.

Analytic Trends

Data Sources	**Algorithms**
Mobile devices Auto telematics Home sensors (Internet of Things) Drones, micro satellites	Statistical learning Artificial intelligence Structural models
Data	**Software**
Big data (text, speech, image, video) Behavioral data (including social media) Credit, trading, financial data	Text analysis, semantics Voice recognition Image recognition Video recognition

Analytic Trends

Data Sources	**Algorithms**
Mobile devices Auto telematics Home sensors (Internet of Things) Drones, micro satellites	Statistical learning Artificial intelligence Structural models
Data	**Software**
Big data (text, speech, image, video) Behavioral data (including social media) Credit, trading, financial data	Text analysis, semantics Voice recognition Image recognition Video recognition

Analytics is now being used in almost every facet of insurance operations including not only marketing of products but also pricing, underwriting, claims management, and reserving.

Insurance Data Are Complicated

- Unlike other products/services, insurance data are complicated because the cost is unknown when the contract is agreed upon
- Insurance analytics began with a focus on personal general (non-life) insurance such as auto and homeowners
  - An early example was the use of credit scores by Progressive Insurance for automobile insurance.
  - This is a competitive marketplace where segmentation makes a difference in market share

Insurance Data Are Complicated

- Unlike other products/services, insurance data are complicated because the cost is unknown when the contract is agreed upon
- Insurance analytics began with a focus on personal general (non-life) insurance such as auto and homeowners
  - An early example was the use of credit scores by Progressive Insurance for automobile insurance.
  - This is a competitive marketplace where segmentation makes a difference in market share
- Receiving less attention (so far) are *liability* lines
  - Possibly because claims take longer to develop, suggesting a need for different modeling techniques
  - This long development time also impacts use of analytic techniques in life/retirement systems lines of business
- Due to the success of analytics in personal lines, many are looking to import these approaches into *commercial* lines. More difficult here due in part to the heterogeneity of risks.

Personal General Insurance Data

- Personal, general insurance data exhibit complex features
  - Insurance data typically has lots of explanatory variables. Lots.
    - It is common to handle this aspect using regression techniques

Personal General Insurance Data

- Personal, general insurance data exhibit complex features
  - Insurance data typically has lots of explanatory variables. Lots.
    - It is common to handle this aspect using regression techniques
  - Loss distributions are typically skewed and heavy-tailed

Personal General Insurance Data

- Personal, general insurance data exhibit complex features
  - Insurance data typically has lots of explanatory variables. Lots.
    - It is common to handle this aspect using regression techniques
  - Loss distributions are typically skewed and heavy-tailed
  - It is common to encounter situations where there are non-continuous discrete mass points in the distribution
    - Two-Part? For example, loss or no loss.
    - Losses may be censored by policy limits
    - Losses may be truncated by deductibles

Personal General Insurance Data

- Personal, general insurance data exhibit complex features
  - Insurance data typically has lots of explanatory variables. Lots.
    - It is common to handle this aspect using regression techniques
  - Loss distributions are typically skewed and heavy-tailed
  - It is common to encounter situations where there are non-continuous discrete mass points in the distribution
    - Two-Part? For example, loss or no loss.
    - Losses may be censored by policy limits
    - Losses may be truncated by deductibles
- Although complex, these features are routine
  - Insurance has provided a fertile field to develop specialized approaches to handle these features
  - We now know how to handle these features using modern statistical approaches

Illustrative Examples

- Spanish automobile and homeowners data
  - An example of personal general insurance
- Wisconsin local government property insurance data
  - An example of a (small) commercial line insurance
- Medical malpractice data
  - An example of a liability line of insurance

Spanish Auto and Home Data

- We examine **longitudinal data** from a major Spanish insurance company that offers automobile and homeowners insurance.
- As in many countries, in Spain
  - vehicle owners are obliged to have some minimum form of insurance coverage for personal injury to third parties.
  - Homeowners insurance, on the other hand, is optional.
- The dataset tracks 40,284 clients over five years, between 2010 and 2015, who subscribed to both automobile and homeowners insurance.
- Observations include customer characteristics such as age, gender, vehicle characteristics for auto insurance, information on property for homeowners, records of claims, and renewal information e.g., date of renewal.

Number of policies and lapse by year

	2010	2011	2012	2013	2014
Number of customers at the beginning of the period	40284	29818	22505	17044	13284
Number of customers that cancel at least one policy per period	10466	7313	5461	3760	2296
Rate of non-renewals for at least one policy (%)	26%	25%	24%	22%	17%

Claim Summary Statistics

	2010	2011	2012	2013	2014
Auto					
Clients with positive Claims	769	547	318	209	124
Average Number of Claim	0.04	0.03	0.03	0.02	0.02
Average cost of Claim	1539.99	1689.84	2031.2	1629.18	1222.13
Home					
Clients with positive Claims	660	531	448	310	240
Average Number of Claim	0.03	0.03	0.04	0.03	0.03
Average cost of Claim	447.85	501.59	410.73	348.1	508.86

Many Rating Factors Available

Variable	Description
year	
gender	1 for male, 0 for female
Age_client	age of the customer
Client_Seniority	the number of years with the company
metro_code	1 for urban or metropolitan, 0 for rural
Car_power_M	power of the car
Car_2ndDriver_M	presence of a second driver
Policy_PaymentMethodA	1 for annual payment, 0 for monthly payment
Insuredcapital_continent_re	value of the property
appartment	1 for apartment, 0 for houses or semi-attached houses
Policy_PaymentMethodH	1 for annual payment, 0 for monthly payment

Questions

- What is the relationship among lapsation and claims outcomes after controlling for rating factors?
  - We introduce a *copula* regression model to represent this relationship
- It turns out that lapse is related to claims. There are also other variables where there are societal concerns (gender for this study). Can (and should) we use this information in pricing?

Local Government Property Insurance Fund

- Was established to provide property insurance for local government entities that include counties, cities, towns, villages, school districts, and library boards.
- Covers over a thousand local government entities who pay approximately $25 million in premiums each year and receive insurance coverage of about $75 billion.
- Offers three major groups of insurance coverage: building and contents (BC), inland marine (construction equipment), and motor vehicles.
- Acts as a stand-alone insurance company, charging premiums to each local government entity (policyholder) and paying claims when appropriate.
- Not permitted to deny coverage for local government entities. Thus, the LGPIF acts as a “residual” market to a certain extent

Building and Contents Claims Summary

- Number of policyholders is declining yet coverage is steady

Year	Average Frequency	Average Severity	Average Coverage	Number of Policyholders
2006	1.015	17,729	32,498,186	1,154
2007	1.235	15,158	35,275,949	1,138
2008	1.041	10,728	37,267,485	1,125
2009	1.277	9,934	40,355,382	1,112
2010	1.285	33,026	41,242,070	1,110
2011	1.036	20,554	42,503,989	1,094

- Going forward, 2006–2010 is for training, 2011 is for validation

Description of Base Rating Variables

Variable	Description
`EntityType`	Categorical variable that is one of six types: (Village, City, County, Misc, School, or Town)
`LnCoverage`	Total building and content coverage, in logarithmic millions of dollars
`LnDeduct`	Deductible, in logarithmic dollars
`NoClaimCredit`	Binary variable to indicate no claims in the past two years
`Fire5`	Binary variable to indicate the fire class is below 5 (The range of fire class is $0 \sim 10$)

Claims Summary by Entity Type, Fire Class, and No Claim Credit

Variable	Number of Policies	Average Frequency	Average Claim
`EntityType`			
Village	1,341	0.529	11,869.75
City	793	2.042	39,177.27
County	328	4.973	95,832.87
Misc	609	0.204	40,011.89
School	1,597	1.500	70,606.31
Town	971	0.118	18,449.46
`Fire5`-No	2,508	0.563	18,346.54
`Fire5`–Yes	3,131	1.655	68,798.51
`NoClaimCredit`–No	3,786	1.571	53,283.61
`NoClaimCredit`–Yes	1,853	0.349	32,666.54
Total	5,639	1.169	49,358.53

- There is substantial variation in the claims distribution by each rating variable
- By itself, `Fire5` is counter-intuitive. We anticipate lower claims when the fire class is below 5 (`Fire5`=Yes)

Description of Endorsements

Variable	Description
Business Interruption	Reimburses an insured for business interruption (lost profits and continuing fixed expenses)
Accounts Receivable	Adds coverage for money owed by its debtors during business interruption due to a covered loss.
Pier and Wharf	Loss of watercraft, by the pressure of ice or water on piers and wharves
Fine Arts	Adds coverage (agreed value) on fine arts, either per item or per exhibit
Golf Course Grounds	Adds coverage to golf course type property such as greens, tees, fairways, etc.
Special Use Animal	Adds coverage for police enforcement animals, such as dogs and horses
Zoo Animals	Adds coverage for zoo animals. Animal mortality is specifically excluded.
Vacancy Permit	Allows claims from covered losses arising from vacant property
Monies and Securities	Adds coverage for monies and securities for loss by theft, disappearance, or destruction (A: loss inside premise, B: loss outside premise).
Other Endorsements	Other additional endorsements, including ordinance & law, and extra expenses

Questions

- This is a rich, microlevel database, a core of which is publicly available
- The usual impediments for making data publicly available for research are:
  - Individual privacy
  - Time and expense to the supporting firm
  - Competition considerations
- We have use this database to study endorsements, dependence modeling, lapsation, and other aspects in a series of publications.
- You are welcome to use this data, as well.

More information at my website:
https://sites.google.com/a/wisc.edu/jed-frees/

Analytics and Medical Malpractice

- Medical malpractice, *a.k.a.* medical professional liability, is a type of insurance that provides compensation to an injured patient and families due to healthcare provider negligence
- This began as an insurance product in the late 1800's, so why do we need analytics now?

Analytics and Medical Malpractice

- Medical malpractice, *a.k.a.* medical professional liability, is a type of insurance that provides compensation to an injured patient and families due to healthcare provider negligence
- This began as an insurance product in the late 1800's, so why do we need analytics now?
  - This line of business has small number of claims that are highly volatile $\Rightarrow$ need non-traditional analytic tools
  - This line of business is changing
    - New court cases
    - Infusion of electronic medical record systems
    - A changing business $\Rightarrow$ need non-traditional analytic tools

Regulatory Restrictions

- Regulations that might restrict use of variables primarily come through the
  - marketing,
  - underwriting (initial and renewal),
  - claims, and
  - ratemaking functions
- For internal company management, e.g., loss reserving and solvency, less of an issue.
  - Certainly not an issue with e.g. risk based capital rules that are formula driven
  - Might need to be careful about ORSA (Own Risk and Solvency Assessment) as these are company specific.

Pricing Principles

- In most marketplaces, prices for products are dictated by forces of supply and demand.
- Not so in personal insurance - pricing constraints placed by regulators.
- Pricing restrictions more common in US than, e.g., EU
  - In EU, gender is not a permitted variable
  - In the US, lapse may not be permitted (varies by state)
- Fewer restrictions in commercial insurance
  - From a regulatory standpoint, this is appropriate when the purchaser of insurance has much less knowledge than the insurer - asymmetric information.

Pricing Restrictions

- Regulators do allow for some discrimination in insurance, not everyone is forced to pay the same price.
- Examples of (potentially) prohibited rating factors
  - Ethnicity/Race
  - Political affiliation
  - Religion
  - Hobbies and other interests unrelated to risk of loss
  - Credit scores, Economic status, and on
- Should we collect these variables?
  - Sometimes no. Just the act of collecting them means that we may become biased in subtle ways, a type of moral hazard.
  - Sometimes yes. Either the status is uncertain or we may wish to understand their effects and prove that we are not being biased.

Cost Based Rates

- US. The general principle is that rates are allowed to cover the cost to transfer a risk from a policyholder to the insurer. The traditional actuarial pricing model produces "indicated" rates:

$$Rate = \text{function}(Expected\ Cost, Contract\ Risk\ Premium, Expense\ Cost, Company\ Profit)$$

  - Also known as a "cost-based indication" - an actuarially sound estimate of the cost to transfer covered risk.
- Historically, procedures have been applied that permit a bit of discretion or "actuarial judgement," including:
  - Simplification (e.g., rounding)
  - Consistency (smoothing to reflect a simpler or more logical pattern, capping relationships)
  - Disruption - mitigate changes in rates over time
- In today's world, more variables and algorithms make regulators concerned about "judgement"

Shopping is the Norm in most Markets

- Should we allow prices to depend on demand-side variables?
  - Lapse retention
  - Other insurer prices
  - Market availability of insurance...
- The ability to collect detailed data on risk retention, defecting clients, quote data, and closure rates by numerous risk characteristics provides a wealth of additional information beyond point estimate indications of the cost of risk transfer.
- Many consumer advocates oppose the use of such data.
  - Seems unfair when used in the insurance context - Whether or not someone comparison shops is not a fair predictor of risk.
  - Penalizes loyal customers
    - Who think they are getting a benefit by being loyal and do not know they may be overpaying
    - Who may have bundled to get a better deal but are not getting one

Regulation and Modeling

- Analysts need to be aware of the purposes/goals when developing models for prediction
- With the same data, you could derive **different models** for:
  - marketing,
  - underwriting, and
  - ratemaking.
- For example in our Spanish context, gender is not permissible for ratemaking but it may be helpful to determine a customer lifetime value model (e.g., profitability over several contract periods).
- Lapse is another good example.
  - Lapse differs because a model with and without lapse is not simply adding another variable to a regression equation.
  - *Longitudinal data with informative dropout* is the technical term

Dependence and Foundations of Insurance

- Insurers exist to spread risks
- Insurance systems are predicated on the pooling of contracts
- Insurers pool risks in order to enjoy the benefits of diversification. But those benefits depend upon relationships among risks.
  - In Insurance 101, we teach about the law of averages, implicitly assuming independence among risks

Dependence and Foundations of Insurance

- Insurers exist to spread risks
- Insurance systems are predicated on the pooling of contracts
- Insurers pool risks in order to enjoy the benefits of diversification. But those benefits depend upon relationships among risks.
  - In Insurance 101, we teach about the law of averages, implicitly assuming independence among risks
  - Some risks provide a natural hedge for one another, thus providing diversification benefits.
    - Think of the mortality risk in life insurance and annuities. If policyholders live longer than anticipated, then insurers pay less in life insurance yet more in annuities, and vice-versa.
  - Some risks naturally depend upon one another, such as the risk of flood to homes that are located close to one another. Few diversification benefits to pooling of such risks.

Dependence is Fundamental to Financial Modeling

- Think of a basic concept from financial portfolios

$$Var(X+Y) = Var(X) + Var(Y) + 2 \times Cov(X,Y)$$

- To understand the variability of the portfolio return, $Var(X+Y)$, you need to under the covariance (dependence) between the two risks which is $Cov(X,Y)$

Sources of Dependencies in Insurance Operations

- There can be several sources of dependencies
  - Several coverages within a policy (e.g., auto, each with a set of limits, deductibles, ...)
  - A person can own several (a “bundle” of) contracts
  - Several people under a single contract (e.g., auto)
  - Multi-level contracts, (e.g., geography, industry, plants,...)
  - Spatial dependencies (e.g., hurricanes, weather-related)
  - Time
  - Frequency and severity

How and Why of Modeling Dependencies

- To model dependencies, I advocate the use of a copula. This tool allows us:
  - to preserve work done that explores the complex marginal features
  - to account for a complex array of dependencies
- Until recently, a limitation of this tool has been a technical point on the non-uniqueness in the presence of discreteness
- Modeling of dependencies is most easily motivated through:
  - Maintaining adequate capital
  - Utilizing appropriate risk controls

Copula Regression Modeling

- Consider the following generic situation. There are $i = 1, \ldots, n$ independent risks. For each risk i, we observe p outcomes $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ and covariates $\mathbf{x}_i$.
- Suppressing the i subscript, the joint distribution of the risks is

$$F(y_1, \ldots, y_p) = \Pr(Y_1 \leq y_1, \ldots, Y_p \leq y_p) = C\left(F_1(y_1), \ldots, F_p(y_p)\right).$$

- Special Cases
  - Multivariate Severity - Yang et al. (2011) investigate the three outcomes of interest, bodily injury, liability payments and the time-to-settlement, using auto injury data from the Insurance Research Council's Closed Claim Survey
  - Longitudinal - Frees and Wang (2005) study automobile bodily injury liability claims from a sample of n=29 Massachusetts towns over $m = 6$ years
  - Multivariate Frequency Severity - Czado et al. (2012) fit a Gaussian copula to the number and average claim size for 12,850 claims from auto policies
  - Multivariate Frequencies - Nikoloulopoulos (2013) gives several applications

Insurance Analytics has Unique Features

Data

- Costs not known in advance
- Competitive personal lines marketplaces uses extensive data
- Liability and life lines require careful thinking of dynamic/temporal aspects

Regulation

- Restrictions in many marketplaces on permissable variables for pricing/ratemaking and marketing
- Provide different models for different purposes

Dependence Among Risks

- More is not necessarily better with dependent risks
- Dependence is a core element for financial products, we need to integrate this aspect more into our thinking at the micro-level

Conclusion

As analysts, our role is to provide models and analytics that help sharpen insurance decision-making, thereby improving the marketplace and helping society.

More information at my website:
https://sites.google.com/a/wisc.edu/jed-frees/

Thank you for your kind attention.

Copula Framework

- Think of the world as
  - Y - insurance losses
  - X_P (permitted variables)
  - X_{NP} (non-permitted variables)
- The joint distribution function is $F(Y, X_P, X_{NP})$.
- There exists a copula such that this joint distribution can be written as $C(F_{Y,X_P}(Y, X_P), F_{X_{NP}}(X_{NP}))$.
  - We think about the model $F_{Y,X_P}(Y, X_P)$ as that which is allowed for rating purposes.
  - However, company strategy is based on the joint model - the copula tells about the association between the rating model and the non-permitted variables.

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim
- $\mathbf{x}$ - risk characteristics, e.g., policyholder age

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim
- $\mathbf{x}$ - risk characteristics, e.g., policyholder age
- i - subscript to denote cross-sectional observations, e.g., $i = 1, \ldots, n = 10,000$

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim
- $\mathbf{x}$ - risk characteristics, e.g., policyholder age
- i - subscript to denote cross-sectional observations, e.g., $i = 1, \ldots, n = 10,000$
- t - subscript to denote temporal observations, e.g., $Y_{it}, \mathbf{x}_{it}$, $t = 1, \ldots, m = 5$

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim
- $\mathbf{x}$ - risk characteristics, e.g., policyholder age
- i - subscript to denote cross-sectional observations, e.g., $i = 1, \ldots, n = 10,000$
- t - subscript to denote temporal observations, e.g., $Y_{it}, \mathbf{x}_{it}$, $t = 1, \ldots, m = 5$
- $\mathbf{Y}$ - multivariate risks, e.g., $\mathbf{Y} = (Y_1, Y_2)$ for auto and homeowners claims

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim
- $\mathbf{x}$ - risk characteristics, e.g., policyholder age
- i - subscript to denote cross-sectional observations, e.g., $i = 1, \ldots, n = 10,000$
- t - subscript to denote temporal observations, e.g., $Y_{it}, \mathbf{x}_{it}$, $t = 1, \ldots, m = 5$
- $\mathbf{Y}$ - multivariate risks, e.g., $\mathbf{Y} = (Y_1, Y_2)$ for auto and homeowners claims
- L - binary variable to indicate that the policy has lapsed (conversely, $1 - L$ indicates retention)

Some Notation for Joint Loyalty and Risk

- Y - an insurance risk, e.g., auto claim
- $\mathbf{x}$ - risk characteristics, e.g., policyholder age
- i - subscript to denote cross-sectional observations, e.g., $i = 1, \ldots, n = 10,000$
- t - subscript to denote temporal observations, e.g., $Y_{it}, \mathbf{x}_{it}$, $t = 1, \ldots, m = 5$
- $\mathbf{Y}$ - multivariate risks, e.g., $\mathbf{Y} = (Y_1, Y_2)$ for auto and homeowners claims
- L - binary variable to indicate that the policy has lapsed (conversely, $1 - L$ indicates retention)

Copula Modeling of Lapse and Claims

1. **Copula Regression**. Estimate parameters of a copula regression model for multivariate (insurance) risks
   - Insurance risks are frequency (discrete), severity (continuous), or claim amounts (a hybrid combination of zeros and continuous outcomes) - thus, likelihoods become complex
   - Introduce a generalized method of moment (GMM) technique to handle this complexity
   - Allows us to handle high-dimensional problems even in the presence of discreteness

Copula Modeling of Lapse and Claims

1. **Copula Regression**. Estimate parameters of a copula regression model for multivariate (insurance) risks
   - Insurance risks are frequency (discrete), severity (continuous), or claim amounts (a hybrid combination of zeros and continuous outcomes) - thus, likelihoods become complex
   - Introduce a generalized method of moment (GMM) technique to handle this complexity
   - Allows us to handle high-dimensional problems even in the presence of discreteness
2. **Lapsation**. Extend this general copula regression framework to include lapse
   - This gives a method of handling a censoring mechanism that may depend on the outcomes
   - For example, a driver with a large auto accident claim may be more likely to lapse a policy

Why Lapse may be useful in Pricing

- Natural struggle between the price of an insurance risk and loyalty.
  - The higher the price, higher is the probability to lapse the policy.
  - Companies typically separate the processes of calculating price and renewal prospects.
- Why is a joint model important?
  - Y insurance claim, L is a lapse (=1 if lapse)
  - For existing customers, insurers expect to see revenue EY for that proportion that renew 1-EL. Total revenue is EY (1-EL).
  - The cost is E $[Y(1-L)]$.
  - To see why the dependence matters, write this as

$$\begin{aligned} \mathrm{E}\, Y(1-L) &= \mathrm{E}\, Y - [(\mathrm{E}\, Y)(\mathrm{E}\, L) + \mathrm{Cov}(Y,L)] \\ &= \mathrm{E}\, Y - (\mathrm{E}\, Y)(\mathrm{E}\, L) - \mathrm{Corr}(Y,L)\mathrm{SD}_Y\mathrm{SD}_L \end{aligned}$$

  - The higher the correlation $\mathrm{Corr}(Y,L)$, the lower is the cost.

Dependence Modeling Strategy

1. Develop a dependence structure that adequately captures relationships among risks
2. Identify risk measures that summarize the uncertainty
3. With a summary measure of the uncertainty that depends on risk control mechanisms, risk managers will be able to observe changes in the risk summary under different risk control settings, and thereby make informed decisions.

Our work focuses on (1) and (3)

Risk Controls for Managing Individual Risks

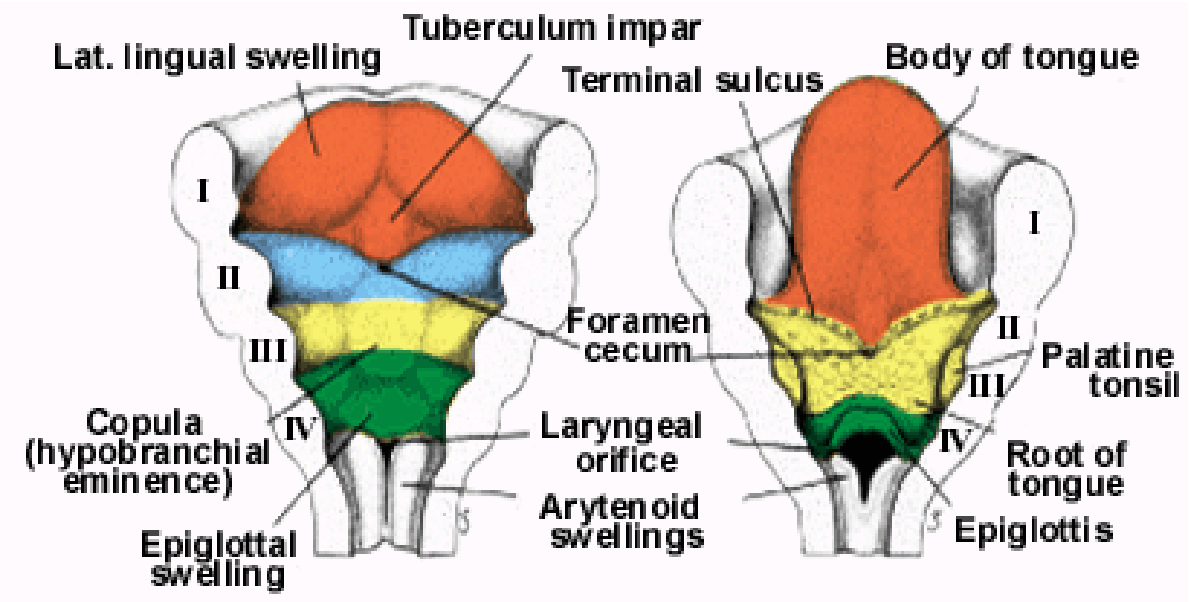
- Academic literature focuses on assessing effects of risk controls when managing **individual** risks (in isolation of other risks).
- *Risk Controls for Managing Individual Risks:*
  - Change the exposure through deductibles, coverage limits, and coinsurance
  - Establish pricing mechanisms at initial acquisition and renewal stages
  - Mitigate risks through reinsurance arrangements
- For the reinsurance, "quota share reinsurance" is a form of proportional reinsurance which specifies that a fixed percentage of each policy written will be transferred to the reinsurer.

Risk Controls for Managing a *Portfolio* of Risks

Risk Control	**Example**
Elect to not insure (or limit the coverage) selected types of perils (a "peril" is a cause of loss)	Earthquakes in homeowners insurance
Elect to not insure (or limit the coverage) selected types of coverages	Not offer comprehensive automobile coverage and write only third party liability in motor insurance
Elect to not underwrite selected groups of policies, either at initial application or at renewal	High risk automobile drivers or or policyholders from a specific geographical area
Mitigate risks through reinsurance arrangements	"Excess of loss" is a type of non-proportional agreement at the portfolio level where the primary insurer is responsible for all losses for a portfolio up to a specified amount (the retention limit) and the reinsurer is responsible for the excess

A Copula Is ...

- The word *copula* derives from the Latin noun for a link or tie that connects two different things.
- In linguistics, a copula (plural: copulae) is a word used to link the subject of a sentence with a predicate (a subject complement or an adverbial).
- In music, the term copula can be thought of as the linking of notes together to form a melody.
- In medicine, the copula is part of the tongue and throat area



Copula Definition

- A *copula* is a multivariate distribution function on $[0,1]^p$ with uniform marginals.
  - Let $U_1, \ldots, U_p$ be p uniform random variables on $(0,1)$. Their distribution function

$$\mathrm{C}(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p)$$

is a copula.

Copula Definition

- A *copula* is a multivariate distribution function on $[0,1]^p$ with uniform marginals.
  - Let $U_1, \ldots, U_p$ be p uniform random variables on $(0,1)$. Their distribution function

    $$\mathrm{C}(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p)$$

    is a copula.
- For applications, consider arbitrary marginal distribution functions $\mathrm{F}_1(y_1), \ldots, \mathrm{F}_p(y_p)$.
- Define a multivariate function using the copula

  $$\mathrm{F}(y_1, \ldots, y_p) = \mathrm{C}(\mathrm{F}_1(y_1), \ldots, \mathrm{F}_p(y_p)).$$

  - F is a multivariate distribution function.

Copula Definition

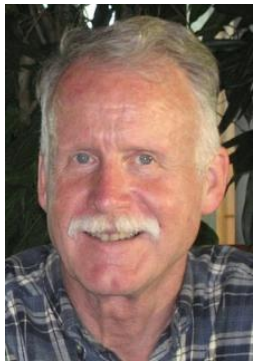
- Let C be a copula – a multivariate distribution function on $[0,1]^p$ with uniform marginals. Then,

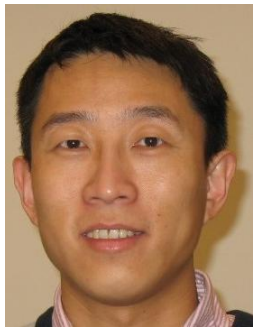
$$\mathrm{F}(y_1,\ldots,y_p) = \mathrm{C}(\mathrm{F}_1(y_1),\ldots,\mathrm{F}_p(y_p)). \tag{1}$$

  is a multivariate distribution function.
- Sklar (1959) established the converse.
  - He showed that *any* multivariate distribution function F can be written using a copula representation.
  - With (1), we can separately infer ideas about the marginals $\mathrm{F}_1(y_1),\ldots,\mathrm{F}_p(y_p)$ and their association mechanism, the copula C.

To Learn More About Copulas

- A probabilistic introduction – Roger Nelsen's *An Introduction to Copulas*
- A statistical introduction – Harry Joe's *Dependence Modeling*
- The `R` package `copula` – Jun Yan and co-authors





Implications for Customer Value

- Pricing models are anchored to the notion that personal lines is a short-term business (contracts typically 6 months or a year)
  - As these are based on costs of insurance, they can be thought of as “supply-side” models
- In contrast, insurers consider non-renewals a critical disruption for their business - losing a customer implies that they will not see benefits in renewal years
(consumers not contractually obligated but likely to renew).
  - Some insurers seek to maximize “lifetime customer value”
  - Marketing groups alter “actuarial” rates to improve customer loyalty
  - Because loyalty is about who buys insurance, it can be thought of as “demand-side” modeling

Single Product, Single Period Price Models

- In an economic sense, an indicated price is based only on the supply side.
  - It does not consider the demand for insurance.
  - How much are consumers willing to pay for insurance?
  - “Willingness to pay” prices are used for informed buyers (no asymmetry problems)
  - In insurance, think of commercial customers.
- Extend the supply-side models to include marketplace demand that depends on existing pools of customers and suppliers.
  - For example, use competitor data as a credibility complement to ensure that the selected rates are reasonable in the marketplace.

Price Optimization

- What is the impact of the rating structure on a firm's portfolio of contracts?
- It is common to examine various business metrics, including
  - performance measures (e.g., combined ratio, total profit) and
  - size measures (e.g., total premium, average premium, policy count).

Multi-Product, Multi-Period World

- **Multi-product**. Consumers need many insurance products simultaneously.
- **Multi-period**. Businesses are built on relationships with customers that extend over many periods.
  - Performance metrics recognize that the firm as a going concern.
- Insurers understand that it is difficult to compete with prices based solely on the costs of risk transfer.
- Ad hoc rate adjustments are made by marketing groups
  - Account for impacts to new business production, close rates, retention, and so forth.
  - The adoption of multiple products – reflected in quantity discounts that acknowledge savings in expenses.
  - Business metrics that measure **change** in the portfolio and performance, such as retention rate, new business closure ratio, new quote count, and new business volume.
  - Another multi-period goal is to “maximize lifetime customer value”
- The impact of the rating scheme on these business metrics will also be affected by the insurers’ underwriting and retention practices.

Multi-Product, Multi-Period Modeling

- Would like a pricing model that integrates demand side and multi-period considerations.
  - Multi-period models that ignore demand side considerations are common in life insurance.
- Formal approaches to developing pricing models that reflect full multi-period marketplace demand has been impeded by the lack of data and the complexity of the pricing decision.
- In contrast, our approach is to:
  - Have one model of cost-based rates
  - Have another model of marketplace demand.
  - Integrate the two using a copula structure.
- Strengths of this approach:
  - We can build on previous experience using cost-based rates.
  - Can explicitly model demand and discuss relationships with costs.
  - Do not need to use demand factors for pricing. This satisfies some regulatory constraints.
  - Not a full endogenous model that explains prices. In this sense, it is more robust in that when behaviors change, certain model components remain stable.

Copula Structure

- Use a copula structure to construct a joint model of variables traditionally thought of as supply or costs and integrate these with demand side variables.
- A joint model provides management with a clearer picture of the world - allows actuaries to give better advice compared to restricting themselves to traditional cost-side variables.
- With a copula representation, we can de-couple different variables. So, if regulatory concerns restrict the use of certain variables, then we can readily accommodate this limitation.

The Copula Approach

- A copula model accounts for relationships among components
  - It accounts for "correlations but not causality."
  - May not be as good as a causal model (in terms of fit, predictions, interpretability) **but**
  - A copula model always exists (Sklar's Theorem)
- The copula approach is comparable the "reduced form" structural models in economics.
- In this work, in our demand side modeling we only consider policyholder retention. We do this in part because
  - Retention represents data that all insurers have in their policyholder database and so our techniques and results will be of interest to a broad audience.
  - Starting with something simple like retention provides a platform for extensions to more complex models where other marketplace variables could be considered (depending on data availability).

Selected References

Czado, Claudia, Rainer Kastenmeier, Eike Christian Brechmann, and Aleksey Min (2012). "A mixed copula model for insurance claims and claim sizes," *Scandinavian Actuarial Journal*, Vol. 2012, pp. 278–305.

Frees, Edward W. and Ping Wang (2005). "Credibility using copulas," *North American Actuarial Journal*, Vol. 9, pp. 31–48.

Ibrahim, Joseph G, Haitao Chu, and Liddy M Chen (2010). "Basic concepts and methods for joint models of longitudinal and survival data," *Journal of Clinical Oncology*, Vol. 28, pp. 2796–2801.

Joe, Harry (2005). "Asymptotic efficiency of the two-stage estimation method for copula-based models," *Journal of Multivariate Analysis*, Vol. 94, pp. 401–419.

——— (2014). *Dependence Modeling with Copulas*. CRC Press.

Liu, Lei (2009). "Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data," *Statistics in Medicine*, Vol. 28, pp. 972–986.

Nikoloulopoulos, Aristidis K (2013). "Copula-based models for multivariate discrete response data," in *Copulae in Mathematical and Quantitative Finance*. Springer, pp. 231–249.

Rizopoulos, Dimitris et al. (2010). "JM: An R package for the joint modelling of longitudinal and time-to-event data," *Journal of Statistical Software*, Vol. 35, pp. 1–33.

Tsiatis, Anastasios A. and Marie Davidian (2004). "Joint modeling of longitudinal and time-to-event data: An overview," *Statistica Sinica*, pp. 809–834.

Yang, Xipei, Edward W. Frees, and Zhengjun Zhang (2011). "A generalized beta copula with applications in modeling multivariate long-tailed data," *Insurance: Mathematics and Economics*, Vol. 49, pp. 265–284.