Predictive Modeling of Insurance Company Operations

Edward W. (Jed) Frees

University of Wisconsin – Madison

May, 2013
Outline

1. Predictive Modeling
2. Two-Part Models
3. Multivariate Regression
4. Multivariate Two-Part Model
5. Gini Index
   - MEPS Model Validation
6. Concluding Remarks
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**
Here are some useful skills/topics

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

As indicated by the title, the focus here is on regression.
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
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
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
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.
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.
- \((\text{Claim recorded})_i = r_i \times y_i\)

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

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

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

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

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:

\[
\ln L = \sum_{i: y_i = d_i} \ln \left( 1 - \Phi \left( \frac{x_i' \beta - d_i}{\sigma} \right) \right) \\
- \frac{1}{2} \sum_{i: y_i > d_i} \left\{ \ln 2\pi \sigma^2 + \frac{(y_i - (x_i' \beta - d_i))^2}{\sigma^2} \right\}
\]

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 \( x_{1i} \) as the set of explanatory variables.
   - Denote the corresponding set of regression coefficients as \( \beta_1 \).
   - 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 \( x_{2i} \) as the set of explanatory variables.
   - Denote the corresponding set of regression coefficients as \( \beta_2 \).
   - Typical models include the linear regression and gamma regression models.
## Full and Reduced Two-Part Models

<table>
<thead>
<tr>
<th>Effect</th>
<th>Full Model</th>
<th>Reduced Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Severity</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.001</td>
<td>0.012</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.395</td>
<td>-0.104</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.108</td>
<td>-0.397</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.008</td>
<td>0.088</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.284</td>
<td>0.078</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>1.958</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.239</td>
<td>1.765</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.132</td>
<td>1.099</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.048</td>
<td>0.356</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>0.002</td>
<td>0.017</td>
</tr>
<tr>
<td>POOR</td>
<td>0.955</td>
<td>4.576</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.087</td>
<td>0.486</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.184</td>
<td>1.422</td>
</tr>
<tr>
<td>VGOOD</td>
<td>0.095</td>
<td>0.736</td>
</tr>
<tr>
<td>MNHPOOR</td>
<td>-0.027</td>
<td>-0.164</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.318</td>
<td>2.941</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.468</td>
<td>-3.131</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.314</td>
<td>-2.318</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.241</td>
<td>-1.626</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.145</td>
<td>-0.716</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.580</td>
<td>4.154</td>
</tr>
<tr>
<td>Scale $\sigma^2$</td>
<td>1.249</td>
<td>1.333</td>
</tr>
</tbody>
</table>

Full and Reduced Two-Part Models
Two-Part Model

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 $F_r(\theta_r)$, where the parameter vector depends on explanatory variables $\theta_r = \theta_r(x)$.

**Part 2.** Similarly, the distribution of $y^*$ can be written as $F_y(\theta_y)$, where $\theta_y = \theta_y(x)$.

- When $\theta_r$ and $\theta_y$ 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.
For an aggregate loss model, we observe $y = (N, S_N)$ and $x$.
- $N$ 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 $(N, y_1^*, \ldots, y_N^*, x)$.
- $y_j^*$ describes the claim amount for each event/episode.
- $S_N = y_1^* + \cdots + y_N^*$ is the aggregate claim amount.
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.
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
See, for example, my 2004 book.
In multivariate analysis, there are several outcomes of interest (multivariate), $y$

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

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

- Use the notation $F_j(\theta_j)$ for the distribution function of $y_j$, $j = 1, \ldots, p$. 

$\text{MEPS Validation}$

$\text{Concluding Remarks}$
Now suppose the outcome of interest is \( y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \).

- Use the notation \( F_j(\theta_j) \) 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 \( \alpha \) is the set of parameters associated with the copula \( C \).
Now suppose the outcome of interest is $y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$.

- Use the notation $F_j(\theta_j)$ 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 $\alpha$ 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 $\theta_j$ depend on explanatory variables $x$ and to use constant $\alpha$. 
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

<table>
<thead>
<tr>
<th></th>
<th>meanSummary</th>
<th>medSummary</th>
<th>sdSummary</th>
<th>minSummary</th>
<th>maxSummary</th>
</tr>
</thead>
<tbody>
<tr>
<td>read</td>
<td>26.6786</td>
<td>25.650</td>
<td>8.753</td>
<td>10.47</td>
<td>43.83</td>
</tr>
<tr>
<td>math</td>
<td>35.8115</td>
<td>34.280</td>
<td>12.091</td>
<td>17.27</td>
<td>66.59</td>
</tr>
<tr>
<td>sci</td>
<td>18.4148</td>
<td>17.920</td>
<td>4.955</td>
<td>9.56</td>
<td>32.88</td>
</tr>
<tr>
<td>ses</td>
<td>-0.0347</td>
<td>-0.037</td>
<td>0.814</td>
<td>-2.41</td>
<td>1.84</td>
</tr>
<tr>
<td>schoolsize</td>
<td>613.3000</td>
<td>500.000</td>
<td>339.158</td>
<td>100.00</td>
<td>1400.00</td>
</tr>
</tbody>
</table>
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.

```r
> 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
    \[
    \frac{\partial^2}{\partial y_1 \partial y_2} F(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} C(F_1(y_1), F_2(y_2)) = f_1(y_1)f_2(y_2)c(F_1(y_1), F_2(y_2)),
    \]

  where \( f_j \) and \( c \) are densities corresponding to the distribution functions \( F_j \) and \( C \).
  - Taking logs, we have
    \[
    L = \ln f_1(y_1) + \ln f_2(y_2) + \ln c(F_1(y_1), F_2(y_2))
    \]

- \( F_1 \) - math - set of beta’s, \( F_2 \) - sci - set of beta’s
- one parameter for the copula
Comparison of Independence to Copula Models

<table>
<thead>
<tr>
<th>Copula Parameter</th>
<th>Coefficient</th>
<th>t-statistic</th>
<th>Coefficient</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.265</td>
<td>22.125</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Math Parameters**

<table>
<thead>
<tr>
<th>Intercept</th>
<th>3.605</th>
<th>137.077</th>
<th>3.611</th>
<th>137.233</th>
</tr>
</thead>
<tbody>
<tr>
<td>minority</td>
<td>-0.074</td>
<td>-3.228</td>
<td>-0.060</td>
<td>-2.635</td>
</tr>
<tr>
<td>ses</td>
<td>0.174</td>
<td>12.980</td>
<td>0.176</td>
<td>13.089</td>
</tr>
<tr>
<td>female</td>
<td>0.015</td>
<td>0.766</td>
<td>0.015</td>
<td>0.795</td>
</tr>
<tr>
<td>public</td>
<td>-0.049</td>
<td>-1.845</td>
<td>-0.056</td>
<td>-2.150</td>
</tr>
<tr>
<td>school1</td>
<td>0.001</td>
<td>0.304</td>
<td>-0.001</td>
<td>-0.192</td>
</tr>
<tr>
<td>urban</td>
<td>0.015</td>
<td>0.620</td>
<td>0.011</td>
<td>0.461</td>
</tr>
<tr>
<td>rural</td>
<td>0.017</td>
<td>0.654</td>
<td>0.031</td>
<td>1.247</td>
</tr>
</tbody>
</table>

**Science Parameters**

<table>
<thead>
<tr>
<th>Intercept</th>
<th>2.985</th>
<th>140.746</th>
<th>2.987</th>
<th>120.422</th>
</tr>
</thead>
<tbody>
<tr>
<td>minority</td>
<td>-0.104</td>
<td>-5.576</td>
<td>-0.089</td>
<td>-4.212</td>
</tr>
<tr>
<td>ses</td>
<td>0.113</td>
<td>10.434</td>
<td>0.110</td>
<td>8.620</td>
</tr>
<tr>
<td>female</td>
<td>-0.024</td>
<td>-1.559</td>
<td>-0.025</td>
<td>-1.421</td>
</tr>
<tr>
<td>public</td>
<td>-0.041</td>
<td>-1.911</td>
<td>-0.039</td>
<td>-1.611</td>
</tr>
<tr>
<td>school1</td>
<td>0.000</td>
<td>-0.128</td>
<td>-0.001</td>
<td>-0.297</td>
</tr>
<tr>
<td>urban</td>
<td>-0.011</td>
<td>-0.557</td>
<td>-0.017</td>
<td>-0.744</td>
</tr>
<tr>
<td>rural</td>
<td>0.007</td>
<td>0.340</td>
<td>0.019</td>
<td>0.789</td>
</tr>
</tbody>
</table>
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)?
Here, we think of $y = (y_1, \ldots, y_T)'$ 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.
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.
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.
You can learn more about copula regression at our Technology Enhanced Learning Project:

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

- 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

\[
\text{Uncertainty} (Z_1 + Z_2) \neq \text{Uncertainty} (Z_2) + \text{Uncertainty} (Z_2).
\]

Need to understand the joint behavior of risks \((Z_1, Z_2)\).
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

\[ \text{Uncertainty}(Z_1 + Z_2) \neq \text{Uncertainty}(Z_2) + \text{Uncertainty}(Z_2) \]

Need to understand the joint behavior of risks \((Z_1, Z_2)\).
Use a multivariate outcome of interest \( y \) where each element of the vector consists of two parts. Thus, we observe

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

and potentially observe

\[
y^* = \begin{pmatrix}
  y_1^* \\
  \vdots \\
  y_p^*
\end{pmatrix}.
\]

- \( r \) - the frequency vector, \( y^* \) as the amount, or severity, vector.
- Decompose the overall likelihood into frequency and severity components

\[
f(r, y^*) = f_1(r) \times f_2(y^* | r)
\]

Let's look at some \textbf{Actuarial Applications}.

- 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

<table>
<thead>
<tr>
<th>Peril ($j$)</th>
<th>Frequency (in percent)</th>
<th>Number of Claims</th>
<th>Median Claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fire</td>
<td>0.310</td>
<td>1,254</td>
<td>4,152</td>
</tr>
<tr>
<td>Lightning</td>
<td>0.527</td>
<td>2,134</td>
<td>899</td>
</tr>
<tr>
<td>Wind</td>
<td>1.226</td>
<td>4,960</td>
<td>1,315</td>
</tr>
<tr>
<td>Hail</td>
<td>0.491</td>
<td>1,985</td>
<td>4,484</td>
</tr>
<tr>
<td>Water Weather</td>
<td>0.776</td>
<td>3,142</td>
<td>1,481</td>
</tr>
<tr>
<td>Water Non-Weather</td>
<td>1.332</td>
<td>5,391</td>
<td>2,167</td>
</tr>
<tr>
<td>Liability</td>
<td>0.187</td>
<td>757</td>
<td>1,000</td>
</tr>
<tr>
<td>Other</td>
<td>0.464</td>
<td>1,877</td>
<td>875</td>
</tr>
<tr>
<td>Theft-Vandalism</td>
<td>0.812</td>
<td>3,287</td>
<td>1,119</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5.889</strong>*</td>
<td><strong>23,834</strong>*</td>
<td><strong>1,661</strong></td>
</tr>
</tbody>
</table>
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.
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

- From the 2006 Medical Expenditure Panel Survey (MEPS), \( n = 18,908 \) individuals
- There are \( p = 5 \) expenditure types

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

* 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
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.

### Counts by Type of Event

<table>
<thead>
<tr>
<th></th>
<th>Singles</th>
<th>Doubles</th>
<th>Triples</th>
<th>Quadruples</th>
</tr>
</thead>
<tbody>
<tr>
<td>OB</td>
<td>6703</td>
<td>OB,OP</td>
<td>1303</td>
<td>OB,OP,ER</td>
</tr>
<tr>
<td>OP</td>
<td>130</td>
<td>OB,ER</td>
<td>917</td>
<td>OB,OP,IP</td>
</tr>
<tr>
<td>ER</td>
<td>396</td>
<td>OB,IP</td>
<td>427</td>
<td>OB,OP,HH</td>
</tr>
<tr>
<td>IP</td>
<td>59</td>
<td>OB,HH</td>
<td>63</td>
<td>OB,ER,IP</td>
</tr>
<tr>
<td>HH</td>
<td>7</td>
<td>OP,ER</td>
<td>30</td>
<td>OB,ER,HH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OP,IP</td>
<td>10</td>
<td>OB,IP,HH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OP,HH</td>
<td>2</td>
<td>OP,ER,IP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ER,IP</td>
<td>44</td>
<td>OP,ER,HH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ER,HH</td>
<td>1</td>
<td>OP,IP,HH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IP,HH</td>
<td>3</td>
<td>ER,IP,HH</td>
</tr>
</tbody>
</table>

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.
Regression Approaches

- **Multinominal Logistic Regressions**
  - With the binary vector $r$, there are $2^p$ possible events

- **Marginal Binary Regressions**
  - With a marginal logistic regression model, we would employ

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

resulting in $\text{logit}(\pi_{i1}) = x_{i1}' \beta_1$.

- Explanatory variables ($x$) and regression coefficients ($\beta$) to depend on the type of outcome.
## Two-Part Models

Multivariate Regression

### Table 13: Marginal Logistic Regressions for Five Types of Events

<table>
<thead>
<tr>
<th>Category Variable</th>
<th>Office-Based Estimate</th>
<th>t-value</th>
<th>Hospital Outpatient Estimate</th>
<th>t-value</th>
<th>Emergency Room Estimate</th>
<th>t-value</th>
<th>Inpatient Estimate</th>
<th>t-value</th>
<th>Home Health Estimate</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demography AGE</td>
<td>0.007</td>
<td>4.046</td>
<td>0.015</td>
<td>6.352</td>
<td>-0.019</td>
<td>-9.295</td>
<td>-0.007</td>
<td>-2.895</td>
<td>0.030</td>
<td>4.415</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.584</td>
<td>16.167</td>
<td>0.398</td>
<td>7.935</td>
<td>0.201</td>
<td>4.388</td>
<td>0.563</td>
<td>8.629</td>
<td>0.246</td>
<td>1.672</td>
</tr>
<tr>
<td>Ethnicity ASIAN</td>
<td>-0.492</td>
<td>-5.853</td>
<td>-0.486</td>
<td>-3.163</td>
<td>-0.621</td>
<td>-3.764</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLACK</td>
<td>-0.438</td>
<td>-9.118</td>
<td>-0.260</td>
<td>-3.717</td>
<td>0.155</td>
<td>2.625</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region NORTHEAST</td>
<td>0.041</td>
<td>0.735</td>
<td>0.598</td>
<td>7.706</td>
<td>0.214</td>
<td>2.748</td>
<td>0.117</td>
<td>1.186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.139</td>
<td>2.733</td>
<td>0.585</td>
<td>8.071</td>
<td>0.310</td>
<td>4.453</td>
<td>0.249</td>
<td>2.825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.044</td>
<td>0.996</td>
<td>0.197</td>
<td>2.855</td>
<td>0.196</td>
<td>3.142</td>
<td>0.197</td>
<td>2.578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Care USC</td>
<td>1.292</td>
<td>32.451</td>
<td>0.822</td>
<td>10.352</td>
<td>0.362</td>
<td>6.020</td>
<td>0.422</td>
<td>5.189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education HIGHSCH</td>
<td>0.104</td>
<td>2.310</td>
<td>0.023</td>
<td>3.590</td>
<td>0.121</td>
<td>1.628</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.289</td>
<td>5.338</td>
<td>0.219</td>
<td>2.979</td>
<td>0.302</td>
<td>3.310</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital Status MARRIED</td>
<td>0.248</td>
<td>5.093</td>
<td>0.209</td>
<td>2.762</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.022</td>
<td>-5.314</td>
</tr>
<tr>
<td>DIVIDSEP</td>
<td>0.144</td>
<td>2.337</td>
<td>0.128</td>
<td>1.505</td>
<td>0.258</td>
<td>4.175</td>
<td></td>
<td></td>
<td>-0.503</td>
<td>-2.675</td>
</tr>
<tr>
<td>Family Size FAMSIZE</td>
<td>-0.114</td>
<td>-9.749</td>
<td>-0.076</td>
<td>-4.990</td>
<td>-0.038</td>
<td>-2.519</td>
<td>0.058</td>
<td>3.201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income HINCOME</td>
<td>0.131</td>
<td>2.059</td>
<td>0.037</td>
<td>-5.148</td>
<td>-0.586</td>
<td>-5.919</td>
<td>-0.229</td>
<td>-1.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MINCOME</td>
<td>0.063</td>
<td>1.084</td>
<td>0.445</td>
<td>-6.404</td>
<td>-0.417</td>
<td>-4.664</td>
<td>-0.518</td>
<td>-2.464</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.072</td>
<td>-1.143</td>
<td>-0.241</td>
<td>-3.183</td>
<td>-0.338</td>
<td>-3.486</td>
<td>-0.372</td>
<td>-1.730</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Health NPOOR</td>
<td>-0.115</td>
<td>-1.363</td>
<td>-0.093</td>
<td>-0.927</td>
<td>-0.080</td>
<td>-0.650</td>
<td>-0.284</td>
<td>-1.063</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIR</td>
<td>1.116</td>
<td>15.785</td>
<td>0.863</td>
<td>9.638</td>
<td>0.850</td>
<td>9.826</td>
<td>0.706</td>
<td>6.438</td>
<td>1.294</td>
<td>3.630</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.587</td>
<td>12.507</td>
<td>0.488</td>
<td>6.421</td>
<td>0.425</td>
<td>5.971</td>
<td>0.409</td>
<td>4.355</td>
<td>0.550</td>
<td>1.522</td>
</tr>
<tr>
<td>VGGOOD</td>
<td>0.421</td>
<td>9.413</td>
<td>0.262</td>
<td>3.443</td>
<td>0.224</td>
<td>3.123</td>
<td>0.228</td>
<td>2.407</td>
<td>0.663</td>
<td>1.799</td>
</tr>
<tr>
<td>Mental Health MNHPoor</td>
<td>0.301</td>
<td>3.792</td>
<td>0.125</td>
<td>1.608</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limitation ANYLIMIT</td>
<td>0.777</td>
<td>15.741</td>
<td>0.726</td>
<td>12.652</td>
<td>0.574</td>
<td>9.918</td>
<td>0.631</td>
<td>8.617</td>
<td>2.051</td>
<td>9.032</td>
</tr>
<tr>
<td>Unemployment UNEMPLOYED</td>
<td>0.139</td>
<td>3.130</td>
<td>0.826</td>
<td>3.804</td>
<td>0.826</td>
<td>3.804</td>
<td>0.826</td>
<td>3.804</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry EDUCHEALTH</td>
<td>0.105</td>
<td>2.021</td>
<td>0.160</td>
<td>1.702</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.745</td>
<td>2.602</td>
</tr>
<tr>
<td>PUBLADMN</td>
<td>0.028</td>
<td>2.251</td>
<td>0.336</td>
<td>2.646</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NATRESOURCE</td>
<td>0.020</td>
<td>2.251</td>
<td>-0.723</td>
<td>-2.300</td>
<td>-0.829</td>
<td>-1.621</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance INSURED</td>
<td>0.672</td>
<td>13.202</td>
<td>0.662</td>
<td>7.949</td>
<td>0.297</td>
<td>5.027</td>
<td>0.532</td>
<td>6.673</td>
<td>1.424</td>
<td>5.364</td>
</tr>
<tr>
<td>Managed Care MANAGEDCARE</td>
<td>0.157</td>
<td>3.540</td>
<td>0.158</td>
<td>2.371</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model fit indices AIC</td>
<td>20,734.74</td>
<td>11,785.83</td>
<td>12,989.63</td>
<td>8,834.53</td>
<td>1,851.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Log-Likelihood</td>
<td>-10,338.37</td>
<td>-5,871.92</td>
<td>-6,471.82</td>
<td>-4,393.27</td>
<td>-908.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Correlations are linear measures – they do not capture associations well for binary data.

One association measure for binary variables is the dependence ratio, \( \tau_{12} \),

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

the ratio of the joint probability to the product of the marginal probabilities.

- In the case of independence, \( \tau_{12} \) to be 1.
- Values of \( \tau_{12} > 1 \) indicate positive dependence.
- Values of \( \tau_{12} < 1 \) indicate negative dependence.
### Joint Counts Among Event Types

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

- Office-Based (OB) expenditures are the most prevalent
- Home Health (HH) expenditures are the least prevalent
### Dependence Ratios Among Event Types

<table>
<thead>
<tr>
<th></th>
<th>OB</th>
<th>OP</th>
<th>ER</th>
<th>IP</th>
<th>HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office-Based (OB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital Outpatient (OP)</td>
<td>1.645</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency Room (ER)</td>
<td></td>
<td>1.963</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inpatient (IP)</td>
<td>1.626</td>
<td>2.499</td>
<td>3.887</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home Health (HH)</td>
<td>1.681</td>
<td>3.198</td>
<td>2.618</td>
<td>6.670</td>
<td></td>
</tr>
</tbody>
</table>

- 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

\[
\text{odds}(r_1) = \frac{\pi_1}{1 - \pi_1} = \frac{\Pr(r_1 = 1)}{\Pr(r_1 = 0)}.
\]

- The odds ratio between \( r_2 \) and \( r_1 \) is

\[
\text{OR}(r_2, r_1) = \frac{\text{odds}(r_2|r_1 = 1)}{\text{odds}(r_2|r_1 = 0)} = \frac{\Pr(r_2 = 1, r_1 = 1) \Pr(r_2 = 0, r_1 = 0)}{\Pr(r_2 = 1, r_1 = 0) \Pr(r_2 = 0, r_1 = 1)}.
\]

- 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

<table>
<thead>
<tr>
<th></th>
<th>OB</th>
<th>OP</th>
<th>ER</th>
<th>IP</th>
<th>HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office-Based (OB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital Outpatient (OP)</td>
<td>10.447</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency Room (ER)</td>
<td>3.371</td>
<td>2.627</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inpatient (IP)</td>
<td>8.454</td>
<td>3.551</td>
<td>8.482</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home Health (HH)</td>
<td>11.902</td>
<td>4.609</td>
<td>3.442</td>
<td>12.717</td>
<td></td>
</tr>
</tbody>
</table>

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 do not account for the association among types.

### Table: Association Test Statistics From Logistic Regression Fits

<table>
<thead>
<tr>
<th></th>
<th>OB</th>
<th>OP</th>
<th>ER</th>
<th>IP</th>
<th>HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office-Based (OB)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hospital Outpatient (OP)</td>
<td>10.735</td>
<td>8.444</td>
<td>9.979</td>
<td>10.072</td>
<td>10.072</td>
</tr>
<tr>
<td>Emergency Room (ER)</td>
<td>9.124</td>
<td>8.444</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Inpatient (IP)</td>
<td>10.313</td>
<td>9.979</td>
<td>25.471</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Home Health (HH)</td>
<td>1.943</td>
<td>2.758</td>
<td>1.763</td>
<td>10.072</td>
<td>-</td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th></th>
<th>Empirical Estimate without Covariates</th>
<th>Likelihood Estimates with Covariates</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bivariate</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OB, OP</td>
<td>10.447</td>
<td>5.603</td>
<td>9.790</td>
</tr>
<tr>
<td>OB, ER</td>
<td>3.371</td>
<td>2.905</td>
<td>10.741</td>
</tr>
<tr>
<td>OB, IP</td>
<td>8.454</td>
<td>6.669</td>
<td>8.378</td>
</tr>
<tr>
<td>OP, ER</td>
<td>2.627</td>
<td>1.985</td>
<td>8.059</td>
</tr>
<tr>
<td>OP, IP</td>
<td>3.551</td>
<td>2.532</td>
<td>8.610</td>
</tr>
<tr>
<td>ER, IP</td>
<td>8.482</td>
<td>6.444</td>
<td>13.186</td>
</tr>
<tr>
<td><strong>Triple</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OB, OP, ER</td>
<td>0.357</td>
<td>0.438</td>
<td>-6.106</td>
</tr>
<tr>
<td>OB, OP, IP</td>
<td>0.218</td>
<td>0.324</td>
<td>-7.443</td>
</tr>
<tr>
<td>OB, ER, IP</td>
<td>0.437</td>
<td>0.397</td>
<td>-7.149</td>
</tr>
<tr>
<td>OP, ER, IP</td>
<td>0.500</td>
<td>0.560</td>
<td>-5.297</td>
</tr>
<tr>
<td><strong>Quadruple</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OB, OP, ER, IP</td>
<td>2.916</td>
<td>2.075</td>
<td>0.851</td>
</tr>
</tbody>
</table>
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??
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
1. Introduction to Predictive Modeling in Actuarial Science | Jed Frees, Glenn Meyers, Richard Derrig
2. Multiple Linear Regression | Margie Rosenberg, Jim Guszczak
3. Regression with Categorical Dependent Variables | Montserrat 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
Predictive Modeling Applications in Actuarial Science

Table of Contents

1. Introduction to Predictive Modeling in Actuarial Science

Fundamentals of Cross-Sectional Regression Modeling

2. Multiple Linear Regression
3. Regression with Categorical Dependent Variables
4. Regression with Count Dependent Variables
5. Generalized Linear Models
6. Frequency/Severity Models

Extended Cross-Sectional Regression Modeling

7. Mixed Models
8. Generalized Additive Models, including Non-Parametric Regression
9. Fat-Tail Regression Models
10. Spatial Statistics
11. Supervised versus Unsupervised Learning
12. Bootstrapping, including Simulation

Bayesian Modeling

13. Introduction to Bayesian Computational Methods
14. Bayesian Regression Models

Longitudinal Modeling

15. Time Series, including Lee-Carter forecasting
16. Longitudinal and Panel Data Models
17. Credibility and Regression Modeling
18. Survival Models, including Cox Regression
19. Claims Triangles/Loss Reserves
20. Transition Modeling
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
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
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?
The relativity is \( R(x_i) = \frac{S(x_i)}{P(x_i)} \).

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.
The relativity is \( R(x_i) = S(x_i)/P(x_i) \).

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.
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
  - $x_i$ - explanatory variables, $P(x_i)$ - premium, $y_i$ - loss, $R_i = R(x_i)$, $I(\cdot)$ - indicator function, and $E(\cdot)$ - mathematical expectation

- The Ordered Lorenz Curve
  - Vertical axis
    
    $$ F_L(s) = \frac{E[yI(R \leq s)]}{E y} = \frac{\sum_{i=1}^{n} y_i I(R_i \leq s)}{\sum_{i=1}^{n} y_i} $$

    the proportion of losses.
  - Horizontal axis
    
    $$ F_P(s) = \frac{E[P(x)I(R \leq s)]}{E P(x)} = \frac{\sum_{i=1}^{n} P(x_i)I(R_i \leq s)}{\sum_{i=1}^{n} P(x_i)} $$

    the proportion of premiums.
Homeowners Example

- 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**

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{F}_P(R_i) - \hat{F}_L(R_i)) \approx \frac{\hat{Gini}}{2},
\]

- 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.
We show that

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

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.
We show that

\[ Gini \approx \frac{2}{n} \operatorname{Cov} ((y - P), \text{rank}(R)) \, . \]

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.
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.
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
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 multinominal 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

<table>
<thead>
<tr>
<th></th>
<th>Mean Absolute Error</th>
<th>Mean Absolute Percentage Error</th>
<th>Root Mean Square Error</th>
<th>Pearson</th>
<th>Spearman</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One Part Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BasicOnePart</td>
<td>3,874.938</td>
<td>227.540</td>
<td>13,774.179</td>
<td>25.680</td>
<td>44.688</td>
<td>18.296</td>
</tr>
<tr>
<td>LogOnePart</td>
<td>2,684.980</td>
<td>6,993.282</td>
<td>14,300.376</td>
<td>22.565</td>
<td>51.515</td>
<td>18.608</td>
</tr>
<tr>
<td>SmearOnePart</td>
<td>10,934.435</td>
<td>174.691</td>
<td>35,244.110</td>
<td>22.565</td>
<td>51.515</td>
<td>18.608</td>
</tr>
<tr>
<td>Tweedie</td>
<td>3,589.482</td>
<td>150.285</td>
<td>13,734.769</td>
<td>26.646</td>
<td>49.616</td>
<td>18.896</td>
</tr>
<tr>
<td><strong>Two Part Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPMLogNSev</td>
<td>2,705.836</td>
<td>525.740</td>
<td>14,209.043</td>
<td>27.036</td>
<td>50.521</td>
<td>18.896</td>
</tr>
<tr>
<td>TPMSmearSev</td>
<td>3,630.774</td>
<td>160.093</td>
<td>13,718.385</td>
<td>27.036</td>
<td>50.521</td>
<td>18.896</td>
</tr>
<tr>
<td>TPMGammaSev</td>
<td>3,579.156</td>
<td>156.046</td>
<td>13,720.311</td>
<td>27.091</td>
<td>50.109</td>
<td>18.893</td>
</tr>
<tr>
<td><strong>Multivariate One Part Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INDBasicOnePart</td>
<td>3,874.938</td>
<td>227.540</td>
<td>13,774.179</td>
<td>25.680</td>
<td>44.688</td>
<td>18.296</td>
</tr>
<tr>
<td>INDLogOnePart</td>
<td>2,719.334</td>
<td>9,387.637</td>
<td>14,433.423</td>
<td>22.584</td>
<td>51.474</td>
<td>18.350</td>
</tr>
<tr>
<td>INDOOnePartTweed</td>
<td>2,781.582</td>
<td>9,787.627</td>
<td>14,513.104</td>
<td>21.256</td>
<td>47.050</td>
<td>18.376</td>
</tr>
<tr>
<td>INDBasicOnePartReduced</td>
<td>3,863.703</td>
<td>211.227</td>
<td>13,773.980</td>
<td>25.694</td>
<td>45.006</td>
<td>18.399</td>
</tr>
<tr>
<td>INDLogOnePartReduced</td>
<td>2,719.343</td>
<td>9,415.450</td>
<td>14,433.430</td>
<td>22.583</td>
<td>51.474</td>
<td>18.353</td>
</tr>
<tr>
<td>INDOOnePartTweedReduced</td>
<td>2,781.683</td>
<td>9,526.059</td>
<td>14,513.085</td>
<td>22.570</td>
<td>46.912</td>
<td>18.509</td>
</tr>
<tr>
<td><strong>Multivariate Two Part Models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INDBasicTPM</td>
<td>3,603.476</td>
<td>265.246</td>
<td>13,731.740</td>
<td>26.811</td>
<td>48.913</td>
<td>18.940</td>
</tr>
<tr>
<td>INDTMLogNSev</td>
<td>2,851.499</td>
<td>258.909</td>
<td>14,075.163</td>
<td>26.351</td>
<td>49.191</td>
<td>18.710</td>
</tr>
<tr>
<td>INDTMGammaSev</td>
<td>3,579.410</td>
<td>146.869</td>
<td>13,736.173</td>
<td>26.677</td>
<td>49.794</td>
<td>18.974</td>
</tr>
<tr>
<td>INDTMLogNSevReduced</td>
<td>2,848.996</td>
<td>258.562</td>
<td>14,062.045</td>
<td>26.879</td>
<td>49.856</td>
<td>18.743</td>
</tr>
<tr>
<td>INDTMGammaSevReduced</td>
<td>3,574.799</td>
<td>149.014</td>
<td>13,726.846</td>
<td>26.883</td>
<td>50.184</td>
<td>18.956</td>
</tr>
<tr>
<td>CellTPMLogNSev</td>
<td>3,053.903</td>
<td>185.672</td>
<td>14,013.866</td>
<td>25.429</td>
<td>50.036</td>
<td>18.773</td>
</tr>
<tr>
<td>DepTPMLogNSevReduced</td>
<td>2,822.43</td>
<td>272.728</td>
<td>14,082.72</td>
<td>26.968</td>
<td>50.058</td>
<td>18.856</td>
</tr>
<tr>
<td>DepTPMGammaSevReduced</td>
<td>3,521.059</td>
<td>152.423</td>
<td>13,732.81</td>
<td>26.784</td>
<td>50.232</td>
<td>18.997</td>
</tr>
</tbody>
</table>
These measures only look at the forecast at a certain point, not the *distribution* of forecasts.

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

Using simulation, it is straightforward 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.
These measures only look at the forecast at a certain point, not the *distribution* of forecasts.

Using simulation, it is straightforward 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.
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.

![Dependence Model Predictions](chart.png)

- Density
- Actual 2007 expenditure
- Dependence Model Predictions
- 0 20 40 60 80 100
- 0.0 0.1 0.2 0.3 0.4
- 0.0
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 $100(1 - \alpha)$ percentile of the distribution.

<table>
<thead>
<tr>
<th>Risk Measure</th>
<th>Percentile 0.50</th>
<th>Percentile 0.75</th>
<th>Percentile 0.90</th>
<th>Percentile 0.95</th>
<th>Percentile 0.98</th>
<th>Percentile 0.99</th>
<th>Percentile 1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>$VaR_{Ind}$</td>
<td>45.05</td>
<td>45.63</td>
<td>46.18</td>
<td>46.56</td>
<td>46.87</td>
<td>47.32</td>
<td>47.44</td>
</tr>
<tr>
<td>$VaR_{Dep}$</td>
<td>41.95</td>
<td>50.51</td>
<td>58.87</td>
<td>65.06</td>
<td>68.99</td>
<td>73.22</td>
<td>79.43</td>
</tr>
<tr>
<td>$CTE_{Ind}$</td>
<td>45.76</td>
<td>46.20</td>
<td>46.67</td>
<td>46.97</td>
<td>47.25</td>
<td>47.55</td>
<td>47.70</td>
</tr>
<tr>
<td>$CTE_{Dep}$</td>
<td>52.40</td>
<td>58.89</td>
<td>66.41</td>
<td>71.31</td>
<td>75.82</td>
<td>81.71</td>
<td>87.11</td>
</tr>
</tbody>
</table>
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.
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.
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 consider “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.