Innovating Healthcare Business Process Service Delivery

Predictive Modeling Basics and Beyond

June 2009

Agenda

1. Background and Issues.
2. Model Objectives.
4. Sample model.
5. Program Planning.
7. Risk Transition.
8. General discussion.
Introductions

Ian Duncan, FSA FIA FCIA MAAA
President, Solucia Consulting, a SCIOinspire Company.

Gary Gau, PhD, ASA
Associate Professor
University of Central Florida

Introduction / Objective

1. What is Predictive Modeling?
2. Types of predictive models.
3. Data and Data Preparation.
Predictive Modeling: A Review of the Basics

Definition of Predictive Modeling

“Predictive modeling is a set of tools used to stratify a population according to its risk of nearly any outcome...ideally, patients are risk-stratified to identify opportunities for intervention before the occurrence of adverse outcomes that result in increased medical costs.”

“Stratified according to risk of event”

"The year 1930, as a whole, should prove at least a fairly good year."

-- Harvard Economic Service, December 1929
**Why do it? Potential Use of Models**

### Program Management Perspective

- Identifying individuals at very high risk of an event (death, LTC, disability, annuity surrender, etc.).
- Identify management opportunities and determine resource allocation/prioritization.

### Identification – how?

- At the heart of predictive modeling!
  - Who?
  - What common characteristics?
  - What are the implications of those characteristics?
- There are many different algorithms for identifying member conditions. THERE IS NO SINGLE AGREED FORMULA.
- Condition identification often requires careful balancing of sensitivity and specificity.
A word about codes and groupers

Codes are the “raw material” of predictive modeling.

Codes are required for payment, so they tend to be reasonably accurate - providers have a vested interest in their accuracy.

Codes define important variables like Diagnosis (ICD-9 or 10); Procedure (CPT); Diagnosis Group (DRG – Hospital); Drug type/dose/manufacturer (NDC); lab test (LOINC); Place of service, type of provider, etc. etc.

“Grouper” models sort-through the raw material and consolidate it into manageable like categories.

Identification – example (Diabetes)

Diabetics can be identified in different ways:

<table>
<thead>
<tr>
<th>Diagnosis type</th>
<th>Reliability</th>
<th>Practicality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician Referral</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Lab tests</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Claims</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Prescription</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Drugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-reported</td>
<td>Low/medium</td>
<td>Low</td>
</tr>
</tbody>
</table>

Medical and Drug Claims are often the most practical method of identifying candidates for predictive modeling.
Identification – example (Diabetes)

Inpatient Hospital Claims - ICD-9 Claims Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>250.xx</td>
<td>Diabetes mellitus</td>
</tr>
<tr>
<td>357.2</td>
<td>Polyneuropathy in diabetes</td>
</tr>
<tr>
<td>362.0, 362.0x</td>
<td>Diabetic retinopathy</td>
</tr>
<tr>
<td>366.41</td>
<td>Diabetic cataract</td>
</tr>
<tr>
<td>648.00-648.04</td>
<td>Diabetes mellitus (as other current condition in mother</td>
</tr>
<tr>
<td></td>
<td>classifiable elsewhere, but complicating pregnancy, childbirth or the</td>
</tr>
<tr>
<td></td>
<td>puerperium.)</td>
</tr>
</tbody>
</table>

Diabetes – additional codes

<table>
<thead>
<tr>
<th>CODES</th>
<th>CODE TYPE</th>
<th>DESCRIPTION - ADDITIONAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0108,</td>
<td>HCPCS</td>
<td>Diabetic outpatient self-management training services, individual or group</td>
</tr>
<tr>
<td>G0109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J1815</td>
<td>HCPCS</td>
<td>Insulin injection, per 5 units</td>
</tr>
<tr>
<td>67227</td>
<td>CPT4</td>
<td>Destruction of extensive or progressive retinopathy, (e.g., diabetic retinopathy) one or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>more sessions, cryotherapy, diathermy</td>
</tr>
<tr>
<td>67228</td>
<td>CPT4</td>
<td>Destruction of extensive or progressive retinopathy, one or more sessions, photocoagulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(laser or xenon arc)</td>
</tr>
<tr>
<td>956.57</td>
<td>ICD-9-CM</td>
<td>Mechanical complications, due to insulin pump</td>
</tr>
<tr>
<td>945.85</td>
<td>ICD-9-CM</td>
<td>Insulin pump status</td>
</tr>
<tr>
<td>953.91</td>
<td>ICD-9-CM</td>
<td>Fitting/adjustment of insulin pump, insulin pump/intubation</td>
</tr>
<tr>
<td>945.46</td>
<td>ICD-9-CM</td>
<td>Encounter for insulin pump training</td>
</tr>
</tbody>
</table>
Diabetes – drug codes

Insulin or Oral Hypoglycemic Agents are often used to identify members. A simple example follows; for more detail, see the HEDIS code-set.

This approach is probably fine for Diabetes, but may not work for other conditions where off-label use is prevalent.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2710</td>
<td>Insulin***</td>
</tr>
<tr>
<td>2720</td>
<td>Sulfonylureas***</td>
</tr>
<tr>
<td>2730</td>
<td>Antidiabetic - Amino Acid Derivatives**</td>
</tr>
<tr>
<td>2740</td>
<td>Biguanides***</td>
</tr>
<tr>
<td>2750</td>
<td>Meglitinide Analogues***</td>
</tr>
<tr>
<td>2760</td>
<td>Diabetic Other***</td>
</tr>
<tr>
<td>2770</td>
<td>Reductase Inhibitors***</td>
</tr>
<tr>
<td>2780</td>
<td>Alpha-Glucosidase Inhibitors**</td>
</tr>
<tr>
<td>2790</td>
<td>Insulin Sensitizing Agents**</td>
</tr>
<tr>
<td>2799</td>
<td>Anti-diabetic Combinations**</td>
</tr>
<tr>
<td></td>
<td>**Oral Anti-Diabetics</td>
</tr>
</tbody>
</table>

More about Grouper Models

Grouper models address several problems inherent in identification from claims (medical and/or drug):

- What “recipe” or algorithm to apply?
- How to keep the algorithm up-to-date?
- How to achieve consistency among users (important, for example, in physician reimbursement or program assessment).

They also have draw-backs:

- Someone else’s definitions;
- Lack of transparency;
- You can’t control sensitivity/specificity trade-off.
Grouper Models – example

<table>
<thead>
<tr>
<th>ICD-9</th>
<th>Dx Group</th>
<th>Condition Category (CC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>402.x</td>
<td>Hypertensive heart disease, with heart failure</td>
<td></td>
</tr>
<tr>
<td>403.1</td>
<td>Hypertensive heart/renal disease, with heart failure</td>
<td></td>
</tr>
<tr>
<td>415.x</td>
<td>Pulmonary vascular disease, except pulmonary embolism</td>
<td></td>
</tr>
<tr>
<td>425.x</td>
<td>Cardiomyopathy/ myocarditis</td>
<td></td>
</tr>
<tr>
<td>428.x</td>
<td>Heart failure</td>
<td></td>
</tr>
</tbody>
</table>

• Each Group and Condition Category becomes an independent variable in a multiple regression equation that results in a weight for that condition;
• Weights correlate with average resource utilization for that condition;
• Some are “trumped” by others (more severe);
• Scores can range from ≅ 0.0 (for young people without diagnoses) to numbers in the 40’s and 50’s (for multiple co-morbid patients).

Construction of a model*

* From Ian Duncan: “Managing and Evaluating Care Management Interventions” (Actex, 2008)

<table>
<thead>
<tr>
<th>Condition Category</th>
<th>Risk Score Contribution</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes with No or Unspecified Complications</td>
<td>0.0</td>
<td>Trumped by Diabetes with Renal Manifestation</td>
</tr>
<tr>
<td>Diabetes with Renal Manifestation</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.0</td>
<td>Trumped by CHF</td>
</tr>
<tr>
<td>Congestive Heart Failure (CHF)</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Drug/Alcohol Dependence</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Age-Sex</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Total Risk Score</td>
<td>4.6</td>
<td></td>
</tr>
</tbody>
</table>
Construction of a model

Grouper/Risk-adjustment theory is that there is a high correlation between risk scores and actual dollars (resources used).

The Society of Actuaries has published three studies that test this correlation. They are available from the SOA and are well worth reading. (See bibliography.) They explain some of the theory of risk-adjusters and their evaluation, as well as showing the correlation between $'s and Risk Scores for a number of commercial models.

Note 1: the SOA tests both Concurrent (retrospective) and Prospective models. Concurrent model correlations tend to be higher.

Note 2: there are some issues with models that you should be aware of:

- They tend to be less accurate at the “extremes” (members with high or low risk scores);
- We have observed an inverse correlation between risk-score and $’s across a wide range of members.

A different approach to grouping

Grouping by Episode

Services related to the underlying diagnosis are grouped

Different diagnosis related groups have different cost weights.

Complete/Incomplete groups

[Diagram showing the timeline of Look-back, Episode 467 Depression, Clean Period, Lab, Prescription, Hospital Admission, Office Visit, Office Visit]
All people are not equally identifiable

Definition Examples:

Narrow: Hospital Inpatient (primary Dx); Face-to-face professional (no X-Ray or Lab)

Broad: Hospital I/P (any Dx); All professional including X-ray, lab.

Rx: Narrow + Outpatient Prescription

Prevalence of 5 Chronic conditions

<table>
<thead>
<tr>
<th></th>
<th>Narrow</th>
<th>Broad</th>
<th>Rx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicare</td>
<td>24.4%</td>
<td>32.8%</td>
<td>30.8%</td>
</tr>
<tr>
<td>Commercial</td>
<td>4.7%</td>
<td>6.3%</td>
<td>6.6%</td>
</tr>
</tbody>
</table>

Identification: False Positives/ False Negatives

False Positive Identification Incidence through Claims
Medicare Advantage Population (with drug benefits)
Diabetes Example

<table>
<thead>
<tr>
<th></th>
<th>Narrow</th>
<th>+ Broad</th>
<th>+ Rx</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1</td>
<td>75.9%</td>
<td>85.5%</td>
<td>92.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Year 2</td>
<td>24.1%</td>
<td>14.5%</td>
<td>7.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Why do it? Model Applications

Reimbursement

- Predicting (normalized) resource use in the population.

Example 1: Normalized resources

Remember the “Scores” we introduced a few slides back?

<table>
<thead>
<tr>
<th>Member Group ID</th>
<th>Condition(s)</th>
<th># members</th>
<th>Score</th>
<th>Risk Total</th>
<th>Expected Cost</th>
<th>Actual Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1080</td>
<td>CHF</td>
<td>2</td>
<td>19.9</td>
<td>39.8</td>
<td>$ 43,780</td>
<td>$ 50,000</td>
</tr>
<tr>
<td>532</td>
<td>Cancer 1</td>
<td>20</td>
<td>8.7</td>
<td>174.2</td>
<td>$ 191,620</td>
<td>$ 150,000</td>
</tr>
<tr>
<td>796</td>
<td>Cancer 2 + Chronic condition</td>
<td>10</td>
<td>16.0</td>
<td>159.7</td>
<td>$ 175,670</td>
<td>$ 160,000</td>
</tr>
<tr>
<td>531</td>
<td>Cancer 2 + No chronic condition</td>
<td>15</td>
<td>9.0</td>
<td>135.3</td>
<td>$ 148,830</td>
<td>$ 170,000</td>
</tr>
<tr>
<td>1221</td>
<td>Multiple chronic conditions</td>
<td>6</td>
<td>4.8</td>
<td>28.8</td>
<td>$ 31,680</td>
<td>$ 50,000</td>
</tr>
<tr>
<td>710</td>
<td>Acute + Chronic Conditions</td>
<td>10</td>
<td>11.1</td>
<td>110.9</td>
<td>$ 121,990</td>
<td>$ 125,000</td>
</tr>
<tr>
<td>882</td>
<td>Diabetes</td>
<td>7</td>
<td>3.7</td>
<td>25.7</td>
<td>$ 28,270</td>
<td>$ 28,000</td>
</tr>
<tr>
<td>967</td>
<td>Cardiac</td>
<td>4</td>
<td>6.1</td>
<td>24.5</td>
<td>$ 26,950</td>
<td>$ 30,000</td>
</tr>
<tr>
<td>881</td>
<td>Asthma</td>
<td>8</td>
<td>3.0</td>
<td>24.1</td>
<td>$ 26,510</td>
<td>$ 40,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>82</td>
<td></td>
<td>723.0</td>
<td>$ 795,300</td>
<td>$ 803,000</td>
</tr>
</tbody>
</table>
Why do it? Model Applications

Program Evaluation

- Predicting resource use based on condition profile.
- Trend Adjustment.

Example 2: Program Evaluation

Typical Program Evaluation Methodology (e.g. DMAA)

Estimated Savings due to reduced PMPY =

Baseline Cost PMPY × Cost Trend = $6,000 × 1.12 = $6,720
Minus: Actual Cost PMPY = $6,300
Equals: Reduced Cost PMPY = $420
Multiplied by: Actual Member Years in Measurement Period = 20,000
Estimated Savings = $8,400,000

Trend can be biased by changes in population risk-profile over time; adjustment for change in average risk will correct for this.
Why do it? Potential Uses of Models

Actuarial, Underwriting
- Calculating new business and renewal premiums

Why do it? Potential Uses of Models

Provider Profiling
- Profiling of provider
- Efficiency Evaluation
- Provider & health plan contracting
Example 4: Provider profiling

Different approaches: provider panel resource prediction (example 1) OR Episode Risk projection

Why do it? Potential Uses of Models

From a Medical Management Perspective

- Identifying individuals at very high risk for high utilization
- Resource allocation and program planning.
Types of Predictive Modeling Tools

- Risk Groupers
- Statistical Models
- Artificial Intelligence

Types of Predictive Modeling Tools

- Risk Groupers
- Statistical Models
- Artificial Intelligence
Uses of Risk Groupers

- Actuarial, Underwriting and Profiling Perspectives
- Medical Management Perspective
- Program Evaluation Perspective

Risk Groupers

What are the different types of risk groupers?
### Selected Risk Groupers

<table>
<thead>
<tr>
<th>Company</th>
<th>Risk Grouper</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHCIS/Ingenix</td>
<td>ERG</td>
<td>Age/Gender, ICD-9 NDC, Lab</td>
</tr>
<tr>
<td>UC San Diego</td>
<td>CDPS</td>
<td>Age/Gender, ICD -9 NDC</td>
</tr>
<tr>
<td>DxCG</td>
<td>DCG RxGroup</td>
<td>Age/Gender, ICD -9 NDC</td>
</tr>
<tr>
<td>Symmetry/Ingenix</td>
<td>ERG PRG</td>
<td>ICD – 9, NDC NDC</td>
</tr>
<tr>
<td>Johns Hopkins</td>
<td>ACG</td>
<td>Age/Gender, ICD – 9</td>
</tr>
</tbody>
</table>

### Risk Grouper Summary

1. Similar performance among all leading risk groupers*.
2. Risk grouper modeling tools use different algorithms to group the source data.
3. Risk groupers use relatively limited data sources (e.g. DCG and Rx Group use ICD-9 and NDC codes but not lab results or HRA information)
4. Most Risk Grouper based Predictive Models combine also use statistical analysis.

* See New SOA study (Winkelman et al) published 2007. Available from SOA.
Types of Predictive Modeling Tools

PM Tools

Risk Groupers

Statistical Models

Artificial Intelligence

Uses of Statistical Models

Statistical models can be used for all 3 uses

Medical Management Perspective

Actuarial, Underwriting and Profiling Perspectives

Program Evaluation Perspective
What are the different types of statistical models?

Types of Statistical Models:
- Logistic Regression
- ANOVA
- Time Series
- Linear Regression
- Non-linear Regression
- Survival Analysis
Types of Predictive Modeling Tools

- Statistical Models
- Risk Groupers
- Artificial Intelligence
What are the different types of artificial intelligence models?
Features of Neural Networks

Reality
- NN tracks complex relationships by resembling the human brain

Perception
- NN can accurately model complicated health care systems

Reality
- Performance equals standard statistical models
- Models overfit data

Neural Network Summary

1. Good academic approach.
2. Few data limitations.
3. Performance comparable to other approaches.
4. Can be hard to understand the output of neural networks (black box).
In Summary

1. Leading predictive modeling tools have similar performance.
2. Selecting a predictive modeling tool should be based on your specific objectives - one size doesn’t fit all.
3. A good predictive model for medical management should be linked to the intervention (e.g. impactibility).
4. “Mixed” models can increase the power of a single model.

Rules vs. Prediction

We are often asked about rules-based models.

1. First, all models ultimately have to be converted to rules in an operational setting.
2. What most people mean by “rules-based models” is actually a “Delphi*” approach. For example, application of “Gaps-in-care” or clinical rules (e.g. ActiveHealth).
3. Rules-based models have their place in Medical Management. One challenge, however, is risk-ranking identified targets, particularly when combined with statistical models.

* Meaning that experts determine the risk factors, rather than statistics.
PM is NOT always about Cost Prediction.....

.....it IS about resource allocation.

- Where/how should you allocate resources?
- Who is intervenable or impactable?
- What can you expect for outcomes?
- How can you manage the key drivers of the economic model for better outcomes?

Cost Stratification of a Large Population

<table>
<thead>
<tr>
<th></th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>67,665</td>
<td>67,665</td>
<td>135,330</td>
<td>676,842</td>
<td>13,537,618</td>
</tr>
<tr>
<td>Actual Cost</td>
<td>$3,204,433,934</td>
<td>$1,419,803,787</td>
<td>$4,624,237,721</td>
<td>$9,680,579,981</td>
<td>$21,973,586,008</td>
</tr>
<tr>
<td>PMPY Total Actual Cost</td>
<td>$47,357</td>
<td>$20,977</td>
<td>$34,170</td>
<td>$14,303</td>
<td>$1,623</td>
</tr>
<tr>
<td>Percentage of Total Cost</td>
<td>14.6%</td>
<td>6.5%</td>
<td>21.1%</td>
<td>44.1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Patients with > $50,000 in Claims

<table>
<thead>
<tr>
<th></th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>19,370</td>
<td>5,249</td>
<td>24,619</td>
<td>32,496</td>
<td>35,150</td>
</tr>
<tr>
<td>Percentage of Total</td>
<td>55.1%</td>
<td>14.9%</td>
<td>70.0%</td>
<td>92.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Cost Stratification of a Large Population

<table>
<thead>
<tr>
<th>Population</th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Cost</td>
<td>67,665</td>
<td>67,665</td>
<td>135,330</td>
<td>676,842</td>
<td>13,537,618</td>
</tr>
<tr>
<td>PMPY Total Actual Cost</td>
<td>$3,204,433,934</td>
<td>$1,419,803,787</td>
<td>$4,624,237,721</td>
<td>$9,680,579,981</td>
<td>$21,973,586,008</td>
</tr>
<tr>
<td>Percentage of Total Cost</td>
<td>14.6%</td>
<td>6.5%</td>
<td>21.1%</td>
<td>44.1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Patients with > $50,000 in Claims

<table>
<thead>
<tr>
<th>Number of Patients</th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of Total</td>
<td>55.1%</td>
<td>14.9%</td>
<td>70.0%</td>
<td>92.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Decreasing Cost / Decreasing Opportunity

Population Risk Ranking

Important Concept: this chart represents Predicted, not Actual Cost.
The Economic Model and Planning a Program

- As the Population Risk Ranking slide shows, all people do not represent equal opportunity.
- The difference in opportunity means that programs need to be well planned.
- It also gives you an opportunity to test the accuracy of different models.

Economic Model: Simple example

- 30,000 eligible members (ee/dep)
- 1,500 - 2,000 with chronic conditions
- 20% “high risk” - 300 to 400
- 60% are reachable and enroll: 180 - 240
- Admissions/high-risk member/year: 0.65
- “Change behavior” of 25% of these:
  - reduced admissions: 29 to 39 annually
  - cost: $8,000/admission
- Gross Savings: $232,000 to $312,000
  - $0.64 to $0.87 pmpm.
Key drivers of the economic model

- Prevalence within the population (numbers)
- Ability to Risk Rank the Population
- Data quality
- Reach/engage ability
- Cost/benefit of interventions
- Timeliness
- Resource productivity
- Random variability in outcomes

Understanding the Economics

[Graph showing DM Program Savings/Costs at different penetration levels]

Penetration (%)

Savings/Cost ($ millions)
Practical Example of Model-Building

What is a model?

- A model is a set of coefficients to be applied to production data in a live environment.

- With individual data, the result is often a predicted value or “score”. For example, the likelihood that an individual will purchase something, or will experience a high-risk event (surrender; claim, etc.).

- For underwriting, we can predict either cost or risk-score.
Available data for creating the score included the following:

- Eligibility/demographics
- Rx claims
- Medical claims

For this project, several data mining techniques were considered: neural net, CHAID decision tree, and regression. The regression was chosen for the following reasons:

With proper data selection and transformation, the regression was very effective, more so than the tree.

1. Split the dataset randomly into halves

Put half of the claimants into an analysis dataset and half into a test dataset. This is to prevent over-fitting. The scoring will be constructed on the analysis dataset and tested on the test dataset. Diagnostic reports are run on each dataset and compared to each other to ensure that the compositions of the datasets are essentially similar. Reports are run on age, sex, cost, as well as disease and Rx markers.
2. Build and Transform independent variables

- In any data-mining project, the output is only as good as the input.
- Most of the time and resources in a data mining project are actually used for variable preparation and evaluation, rather than generation of the actual “recipe”.

3. Build composite dependent variable

- A key step is the choice of dependent variable. What is the best choice?
- A likely candidate is total patient cost in the predictive period. But total cost has disadvantages
  - It includes costs such as injury or maternity that are not generally predictable.
  - It includes costs that are steady and predictable, independent of health status (capitated expenses).
  - It may be affected by plan design or contracts.
- We generally predict total cost (allowed charges) net of random costs and capitated expenses.
- Predicted cost can be converted to a risk-factor.
3. Build and transform Independent Variables

- Typical transforms include:
  - Truncating data ranges to minimize the effects of outliers.
  - Converting values into binary flag variables.
  - Altering the shape of the distribution with a log transform to compare orders of magnitude.
  - Smoothing progression of independent variables.

- A simple way to look at variables:
  - Convert to a discrete variable. Some variables such as number of prescriptions are already discrete. Real-valued variables, such as cost variables, can be grouped into ranges.
  - Each value or range should have a significant portion of the patients.
  - Values or ranges should have an ascending or descending relationship with average value of the composite dependent variable.

Typical "transformed" variable
4. Select Independent Variables

- The following variables were most promising
  - Age - Truncated at 15 and 80
  - Baseline cost
  - Number of comorbid condition truncated at 5
  - MClass
    - Medical claims-only generalization of the comorbidity variable.
    - Composite variable that counts the number of distinct ICD9 ranges for which the claimant has medical claims.
    - Ranges are defined to separate general disease/condition categories.
  - Number of prescriptions truncated at 10

4. Select Independent Variables (contd.)

- Scheduled drug prescriptions truncated at 5
- NClass
  - Rx-only generalization of the comorbidity variable.
  - Composite variable that counts the number of distinct categories distinct ICD9 ranges for which the claimant has claims.
  - Ranges are defined using GPI codes to separate general disease/condition categories.
- Ace inhibitor flag
- Neuroleptic drug flag
- Anticoagulants flag
- Digoxin flag
- Diuretics flag
- Number of corticosteroid drug prescriptions truncated at 2
5. Run Stepwise Linear Regression

An ordinary linear regression is simply a formula for determining a best-possible linear equation describing a dependent variable as a function of the independent variables. But this pre-supposes the selection of a best-possible set of independent variables. How is this best-possible set of independent variables chosen?

One method is a stepwise regression. This is an algorithm that determines both a set of variables and a regression. Variables are selected in order according to their contribution to incremental $R^2$.

5. Run Stepwise Linear Regression (continued)

Stepwise Algorithm

1. Run a single-variable regression for each independent variable. Select the variable that results in the greatest value of $R^2$. This is “Variable 1”.

2. Run a two-variable regression for each remaining independent variable. In each regression, the other independent variable is Variable 1. Select the remaining variable that results in the greatest incremental value of $R^2$. This is “Variable 2.”

3. Run a three-variable regression for each remaining independent variable. In each regression, the other two independent variables are Variables 1 and 2. Select the remaining variable that results in the greatest incremental value of $R^2$. This is “Variable 3.”

......

n. Stop the process when the incremental value of $R^2$ is below some pre-defined threshold.
6. Results - Examples

- Stepwise linear regressions were run using the "promising" independent variables as inputs and the composite dependent variable as an output.
- Separate regressions were run for each patient sex.
- Sample Regressions

  - Female
    - Scheduled drug prescription 358.1
    - NClass 414.5
    - MClass 157.5
    - Baseline cost 0.5
    - Diabetes Dx 1818.9
    - Intercept 18.5

Why are some variables selected while others are omitted? The stepwise algorithm favors variables that are relatively uncorrelated with previously-selected variables. The variables in the selections here are all relatively independent of each other.

---

6. Results - Examples

- Examples of application of the female model

<table>
<thead>
<tr>
<th>Claimant ID</th>
<th>Schedule Drugs</th>
<th>NClass</th>
<th>Cost</th>
<th>Diabetes</th>
<th>MClass</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
<tr>
<td>2</td>
<td>256</td>
<td>4</td>
<td>365</td>
<td>0.5</td>
<td>0.5</td>
<td>315.00</td>
</tr>
<tr>
<td>3</td>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
</tbody>
</table>

Schedule Drugs: Value Range RV < 2 2 < RV < 5 RV > 5
NClass: Value Range RV < 2 2 < RV < 5 RV > 5
Cost: Value Range RV < 2 2 < RV < 5 RV > 5
Diabetes: Value Range RV < 2 2 < RV < 5 RV > 5
MClass: Value Range RV < 2 2 < RV < 5 RV > 5

Transform Function

<table>
<thead>
<tr>
<th>Schedule Drugs</th>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2 &lt; RV &lt; 5</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NClass</th>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2 &lt; RV</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cost</th>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2 &lt; RV</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diabetes</th>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2 &lt; RV</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MClass</th>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2 &lt; RV</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 5</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

TOTAL

<table>
<thead>
<tr>
<th>Schedule Drugs</th>
<th>NClass</th>
<th>Cost</th>
<th>Diabetes</th>
<th>MClass</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
<tr>
<td>256</td>
<td>4</td>
<td>365</td>
<td>0.5</td>
<td>0.5</td>
<td>315.00</td>
</tr>
<tr>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Schedule Drugs</th>
<th>NClass</th>
<th>Cost</th>
<th>Diabetes</th>
<th>MClass</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
<tr>
<td>256</td>
<td>4</td>
<td>365</td>
<td>0.5</td>
<td>0.5</td>
<td>315.00</td>
</tr>
<tr>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Schedule Drugs</th>
<th>NClass</th>
<th>Cost</th>
<th>Diabetes</th>
<th>MClass</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
<tr>
<td>256</td>
<td>4</td>
<td>365</td>
<td>0.5</td>
<td>0.5</td>
<td>315.00</td>
</tr>
<tr>
<td>234</td>
<td>6</td>
<td>434</td>
<td>0.5</td>
<td>0.5</td>
<td>387.25</td>
</tr>
</tbody>
</table>

TOTAL
Evaluation - Case Examples

Background - Case 1

- Large client.
- Several years of data provided for modeling.
- Never able to become comfortable with data which did not perform well according to our benchmark statistics ($/claimant; $pmmp; number of claims per member).

<table>
<thead>
<tr>
<th>BENCHMARK DATA (Commercial only)</th>
<th>PMPM</th>
<th>Claims/Member/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Only</td>
<td>$76.40</td>
<td>14.40</td>
</tr>
<tr>
<td>Rx Only</td>
<td>$16.49</td>
<td>7.70</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$86.89</td>
<td>22.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CLIENT DATA (Commercial includes Capitation)</th>
<th>PMPM</th>
<th>Claims/Member/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical + Rx</td>
<td>$32.95</td>
<td>5.36</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$32.95</td>
<td>5.36</td>
</tr>
</tbody>
</table>
Background - Case 1

- Built models to predict cost in year 2 from year 1.
- Now for the hard part: evaluating the results.

How well does the model perform?

Analysis 1: all groups. This analysis shows that, at the group level, prediction is not particularly accurate, with a significant number of groups at the extremes of the distribution.
How well does the model perform?

Analysis 2: Omitting small groups (under 50 lives) significantly improves the actual/predicted outcomes.

Analysis 3: Weighting the results by the number of lives in the group shows that most predictions lie within +/- 30% of the actual.
**Conclusion**

- Significant data issues were identified and not resolved.
- This was a large group carrier who had many groups “re-classified” during the period. They were unable to provide good data that “matched” re-classified groups to their previous numbers.

- Conclusion: if you are going to do anything in this area, be sure you have good data.

**Background - Case 2.**

- Client uses a manual rate basis for rating small cases. Client believes that case selection/assignment may result in case assignment to rating classes that is not optimal.

- A predictive model may add further accuracy to the class assignment process and enable more accurate rating and underwriting to be done.
Background

- A number of different tree models were built (at client’s request).
- Technically, an optimal model was chosen.

Problem: how to convince Underwriting that:
- Adding the predictive model to the underwriting process produces more accurate results; and
- They need to change their processes to incorporate the predictive model.

Some data

<table>
<thead>
<tr>
<th>Node</th>
<th>PREDICTED Average Profit</th>
<th>PREDICTED Number in Node</th>
<th>PREDICTED Number in Node (Adjusted)</th>
<th>ACTUAL Number in node</th>
<th>ACTUAL Average Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.03)</td>
<td>70</td>
<td>173</td>
<td>170</td>
<td>(0.05)</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>860</td>
<td>2,122</td>
<td>2,430</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>(0.20)</td>
<td>2,080</td>
<td>5,131</td>
<td>6,090</td>
<td>(0.06)</td>
</tr>
<tr>
<td>4</td>
<td>0.09</td>
<td>910</td>
<td>2,245</td>
<td>2,580</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>(0.40)</td>
<td>680</td>
<td>1,678</td>
<td>20</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>(0.27)</td>
<td>350</td>
<td>863</td>
<td>760</td>
<td>0.16</td>
</tr>
<tr>
<td>7</td>
<td>0.11</td>
<td>650</td>
<td>1,604</td>
<td>1,810</td>
<td>0.04</td>
</tr>
<tr>
<td>8</td>
<td>0.53</td>
<td>190</td>
<td>469</td>
<td>470</td>
<td>(0.01)</td>
</tr>
<tr>
<td>9</td>
<td>(0.13)</td>
<td>1,150</td>
<td>2,837</td>
<td>2,910</td>
<td>0.03</td>
</tr>
<tr>
<td>10</td>
<td>0.27</td>
<td>1,360</td>
<td>3,355</td>
<td>3,740</td>
<td>0.04</td>
</tr>
<tr>
<td>11</td>
<td>0.38</td>
<td>1,560</td>
<td>3,849</td>
<td>3,920</td>
<td>(0.07)</td>
</tr>
<tr>
<td>12</td>
<td>0.08</td>
<td>320</td>
<td>789</td>
<td>830</td>
<td>0.08</td>
</tr>
<tr>
<td>13</td>
<td>0.06</td>
<td>12,250</td>
<td>30,221</td>
<td>29,520</td>
<td>0.02</td>
</tr>
<tr>
<td>14</td>
<td>0.27</td>
<td>2,400</td>
<td>5,921</td>
<td>6,410</td>
<td>0.21</td>
</tr>
<tr>
<td>15</td>
<td>(1.07)</td>
<td>540</td>
<td>1,332</td>
<td>1,320</td>
<td>(0.03)</td>
</tr>
<tr>
<td>16</td>
<td>0.07</td>
<td>10,070</td>
<td>24,843</td>
<td>24,950</td>
<td>(0.08)</td>
</tr>
<tr>
<td>17</td>
<td>(0.33)</td>
<td>1,400</td>
<td>3,454</td>
<td>3,250</td>
<td>(0.10)</td>
</tr>
<tr>
<td>18</td>
<td>0.11</td>
<td>4,460</td>
<td>11,003</td>
<td>11,100</td>
<td>0.08</td>
</tr>
<tr>
<td>19</td>
<td>(0.13)</td>
<td>1,010</td>
<td>2,492</td>
<td>2,100</td>
<td>(0.11)</td>
</tr>
</tbody>
</table>

42,310 104,380 104,380 0.005
### How well does the model perform?

<table>
<thead>
<tr>
<th>Node</th>
<th>PREDICTED Average Profit</th>
<th>Predicted Number in Node</th>
<th>Predicted Number in Node (Adjusted)</th>
<th>ACTUAL Number in Node</th>
<th>ACTUAL Average Profit</th>
<th>Directionally Correct (+ or -)</th>
<th>Predicted to be Profitable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.03</td>
<td>70</td>
<td>173</td>
<td>170</td>
<td>0.06</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>860</td>
<td>2,122</td>
<td>2,430</td>
<td>0.07</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>2,080</td>
<td>5,131</td>
<td>6,090</td>
<td>0.06</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>4</td>
<td>0.09</td>
<td>910</td>
<td>2,245</td>
<td>2,580</td>
<td>0.10</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>5</td>
<td>0.40</td>
<td>680</td>
<td>1,678</td>
<td>20</td>
<td>0.02</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>6</td>
<td>0.27</td>
<td>350</td>
<td>863</td>
<td>760</td>
<td>0.16</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>7</td>
<td>0.11</td>
<td>650</td>
<td>1,604</td>
<td>1,810</td>
<td>0.04</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>8</td>
<td>0.53</td>
<td>190</td>
<td>469</td>
<td>470</td>
<td>0.01</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>9</td>
<td>0.13</td>
<td>1,150</td>
<td>2,837</td>
<td>2,910</td>
<td>0.03</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>10</td>
<td>0.27</td>
<td>1,380</td>
<td>3,355</td>
<td>3,740</td>
<td>0.04</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>11</td>
<td>0.38</td>
<td>1,560</td>
<td>3,849</td>
<td>3,920</td>
<td>0.07</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>12</td>
<td>0.08</td>
<td>320</td>
<td>789</td>
<td>830</td>
<td>0.08</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>13</td>
<td>0.06</td>
<td>12,250</td>
<td>30,221</td>
<td>29,520</td>
<td>0.02</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>14</td>
<td>0.27</td>
<td>2,400</td>
<td>5,921</td>
<td>6,410</td>
<td>0.21</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>15</td>
<td>1.07</td>
<td>540</td>
<td>1,332</td>
<td>1,320</td>
<td>0.03</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>16</td>
<td>0.07</td>
<td>10,070</td>
<td>24,843</td>
<td>24,950</td>
<td>0.00</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>17</td>
<td>0.33</td>
<td>1,400</td>
<td>3,454</td>
<td>3,250</td>
<td>0.10</td>
<td>Red</td>
<td>Green</td>
</tr>
<tr>
<td>18</td>
<td>0.11</td>
<td>4,460</td>
<td>11,003</td>
<td>11,100</td>
<td>0.08</td>
<td>Green</td>
<td>Red</td>
</tr>
<tr>
<td>19</td>
<td>0.13</td>
<td>1,010</td>
<td>2,492</td>
<td>2,100</td>
<td>0.11</td>
<td>Red</td>
<td>Green</td>
</tr>
</tbody>
</table>

Predicted to be Profitable (Red = Profitable, Green = Non-profitable)

Total predicted nodes: 6
Total actual nodes: 13
Nodes differing: 11
## Underwriting Decision-making

<table>
<thead>
<tr>
<th>Underwriting Decision</th>
<th>Total Profit</th>
<th>Average Profit per Case</th>
<th>Cases Written</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept all cases as rated.</td>
<td>557.5</td>
<td>0.005</td>
<td>104,380</td>
</tr>
<tr>
<td>Accept all cases predicted to be</td>
<td>1,379.4</td>
<td>0.016</td>
<td>87,760</td>
</tr>
<tr>
<td>profitable; reject all predicted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unprofitable cases.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Underwriting Decision-making

<table>
<thead>
<tr>
<th>Underwriting Decision</th>
<th>Total Profit</th>
<th>Average Profit per Case</th>
<th>Cases Written</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept all cases as rated.</td>
<td>557.5</td>
<td>0.005</td>
<td>104,380</td>
</tr>
<tr>
<td>Accept all cases predicted to be profitable; reject all predicted</td>
<td>1,379.4</td>
<td>0.016</td>
<td>87,760</td>
</tr>
<tr>
<td>unprofitable cases.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept all cases predicted to be profitable; rate all cases</td>
<td>2,219.5</td>
<td>0.021</td>
<td>104,380</td>
</tr>
<tr>
<td>predicted to be unprofitable +10%.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept all cases for which the directional prediction is correct.</td>
<td>2,543.5</td>
<td>0.026</td>
<td>100,620</td>
</tr>
</tbody>
</table>
### Underwriting Decision-making

<table>
<thead>
<tr>
<th>Underwriting Decision</th>
<th>Total Profit</th>
<th>Average Profit per Case</th>
<th>Cases Written</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept all cases as rated.</td>
<td>557.5</td>
<td>0.005</td>
<td>104,380</td>
</tr>
<tr>
<td>Accept all cases predicted to be profitable; reject all predicted unprofitable cases.</td>
<td>1,379.4</td>
<td>0.016</td>
<td>87,760</td>
</tr>
<tr>
<td>Accept all cases predicted to be profitable; rate all cases predicted to be unprofitable +10%.</td>
<td>2,219.5</td>
<td>0.021</td>
<td>104,380</td>
</tr>
<tr>
<td>Accept all cases for which the directional prediction is correct.</td>
<td>2,543.5</td>
<td>0.026</td>
<td>100,620</td>
</tr>
<tr>
<td>Accept all cases for which the directional prediction is correct; rate predicted unprofitable cases by +10%</td>
<td>3,836.5</td>
<td>0.038</td>
<td>100,620</td>
</tr>
<tr>
<td>Accept all cases for which the directional prediction is correct.</td>
<td>2,540.8</td>
<td>0.025</td>
<td>101,090</td>
</tr>
</tbody>
</table>
Example 3: evaluating a high-risk model

Background

- Large health plan client seeking a model to improve case identification for case management.

- Considered two commercially-available models:
  - Version 1: vendor’s typical predictive model based on conditions only. Model is more typically used for risk-adjustment (producing equivalent populations).
  - Version 2: vendor’s high-risk predictive model that predicts the probability of a member having an event in the next 6-12 months.
• Client initially rejected model 2 as not adding sufficient value compared with model 1. (Vendor’s pricing strategy was to charge additional fees for model 2) based on cumulative predictions.
• Looked at over a narrower range, however, the results appear different.
### Analysis

<table>
<thead>
<tr>
<th>Decile From</th>
<th>Decile To</th>
<th>Population Expected</th>
<th>Actual</th>
<th>Predicted Frequency</th>
<th>Actual Frequency</th>
<th>Predictive ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>90%</td>
<td>1,690</td>
<td>808</td>
<td>694</td>
<td>47.8%</td>
<td>41.1%</td>
</tr>
<tr>
<td>90%</td>
<td>80%</td>
<td>1,699</td>
<td>268</td>
<td>321</td>
<td>15.8%</td>
<td>18.9%</td>
</tr>
<tr>
<td>80%</td>
<td>70%</td>
<td>1,657</td>
<td>152</td>
<td>247</td>
<td>9.2%</td>
<td>14.9%</td>
</tr>
<tr>
<td>70%</td>
<td>60%</td>
<td>1,673</td>
<td>107</td>
<td>191</td>
<td>6.4%</td>
<td>11.4%</td>
</tr>
<tr>
<td>60%</td>
<td>50%</td>
<td>1,661</td>
<td>82</td>
<td>168</td>
<td>4.9%</td>
<td>10.0%</td>
</tr>
<tr>
<td>50%</td>
<td>40%</td>
<td>1,760</td>
<td>67</td>
<td>165</td>
<td>3.8%</td>
<td>9.4%</td>
</tr>
<tr>
<td>40%</td>
<td>30%</td>
<td>1,667</td>
<td>50</td>
<td>118</td>
<td>3.0%</td>
<td>7.1%</td>
</tr>
<tr>
<td>30%</td>
<td>20%</td>
<td>1,729</td>
<td>38</td>
<td>92</td>
<td>2.2%</td>
<td>5.3%</td>
</tr>
<tr>
<td>20%</td>
<td>10%</td>
<td>1,624</td>
<td>26</td>
<td>68</td>
<td>1.6%</td>
<td>4.2%</td>
</tr>
<tr>
<td>10%</td>
<td>0%</td>
<td>1,708</td>
<td>91</td>
<td>37</td>
<td>5.3%</td>
<td>2.2%</td>
</tr>
</tbody>
</table>

16,888 | 1,690 | 2,101 | 100% | 124.4% |

---

### Example 4: a wellness model
Using data from a large health plan (multi-million lives; both self-reported data and health claims) we developed a risk-factor model that relates claims dollars to risk factors;

- Multiple regression model;
- 15 different risk factors;
- Multiple categorical responses.

### Solucia Wellness Model

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Variable</th>
<th>Values</th>
<th>Cost Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>1</td>
<td>190</td>
</tr>
<tr>
<td>Personal Disease History 1</td>
<td>Chronic Obstructive Pulmonary Disease (COPD), Congestive Heart Failure (CHF), Coronary Heart Disease (CHD), Peripheral Vascular Disease (PVD) and Stroke</td>
<td>0 (No)</td>
<td>10,553</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Yes)</td>
<td></td>
</tr>
<tr>
<td>Health Screenings</td>
<td>Have you had a SIGMOIDOSCOPY within the last 5 years? (tube inserted in rectum to check for lower intestine problems)</td>
<td>0 (No)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Yes)</td>
<td></td>
</tr>
<tr>
<td>Weight Management</td>
<td>Body Mass Index</td>
<td>26 (Min)</td>
<td>3,069</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40 (Max)</td>
<td>4,722</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45 (Max)</td>
<td>5,312</td>
</tr>
<tr>
<td>Health Screenings</td>
<td>Influenza (flu) within the last 12 months?</td>
<td>0 (No)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Yes)</td>
<td></td>
</tr>
<tr>
<td>Personal Disease History 2</td>
<td>Have you never been diagnosed with any of the following: list of 27 major conditions</td>
<td>0 (No)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Yes)</td>
<td></td>
</tr>
<tr>
<td>Personal Disease History 3</td>
<td>TIA (mini-stroke lasting less than 24 hrs), Heart Attack, Angina, Breast Cancer, Emphysema</td>
<td>0 (No)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Yes)</td>
<td></td>
</tr>
<tr>
<td>Immunizations</td>
<td>Pneumonia</td>
<td>0 (No)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Yes)</td>
<td></td>
</tr>
<tr>
<td>Physical Activity 1</td>
<td>Moderate/Intensity physical activity - minutes per day</td>
<td>0 (Min, no Value)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20 (Max)</td>
<td></td>
</tr>
<tr>
<td>Stress and Well-Being</td>
<td>In the last month, how often have you been angered because of things that happened that were outside your control?</td>
<td>0 (Never, Almost Never)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (Very Often, No Value)</td>
<td></td>
</tr>
</tbody>
</table>

97
### Solucia Wellness Model

#### Skin Protection

Please rate how confident you are that you can have your skin checked by a doctor once a year?

- 1 (Not at all confident) (224)
- 2 (Not confident) (447)
- 3 (Fairly confident) (671)
- 4 (Confident) (894)
- 5 (Very Confident) (1,118)
- 7 (No Value) (1,565)

#### Women's health 1

Are you currently on hormone replacement therapy (Estrogen Therapy, Premarin) or planning to start?

- 0 (No) -
- 1 (Yes) 999

#### Women's health 2

Select the appropriate answer regarding pregnancy status/plan

- 0 (Not Planning (I am planning on becoming pregnant in the next 6 months.)) 590
- 1 (No Value) 1,181
- 3 (Planning (I am planning on becoming pregnant in the next 6 months.)) 1,771
- 4 (Pregnant (I am currently pregnant)) 2,361

#### Physical Activity 2

HIGH intensity activities? (hours per week)

- 0 (Min, No Value) -
- 3 (Max) 617

#### Nutrition

On a typical day, how many servings do you eat of whole grain or enriched bread, cereal, rice, and pasta?

- 0 (None, No Value) -
- 1 (One Three, Four Five) 868
- 2 (SixPlus) 1,736

#### Tobacco

Please rate how confident you are that you can keep from smoking cigarettes when you feel you need a lift.

- 1 (Not at all confident) (294)
- 2.5 (No Value) (441)
- 2 (Not confident) (588)
- 3 (Fairly confident) (883)
- 4 (Confident) (1,177)

---

### Discussion?
Selected references

This is not an exhaustive bibliography. It is only a starting point for explorations.

- Shapiro, A.F. and Jain, L.C. (editors); *Intelligent and Other Computational Techniques in Insurance*; World Scientific Publishing Company; 2003.
- Dove, Henry G., Duncan, Ian, and Robb, Arthur; *A Prediction Model for Targeting Low-Cost, High-Risk Members of Managed Care Organizations*; The American Journal of Managed Care, Vol 9 No 5, 2003
- Berry, Michael J. A. and Linoff, Gordon; *Data Mining Techniques for Marketing, Sales and Customer Support*; John Wiley and Sons, Inc; 2004
- Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G Geoffrey; *Introduction to Linear Regression Analysis*; John Wiley and Sons, Inc; 2001
- Kahneman, Daniel, Slovic, Paul, and Tversky (editors); *Judgment under uncertainty: Heuristics and Biases*; Cambridge University Press; 1982

Selected references (contd.)

Further Questions?

iduncan@soluciaconsulting.com

Solucia Inc.
220 Farmington Avenue, Suite 4
Farmington, CT 06032

860-676-8808

www.soluciaconsulting.com
Introduction to Generalized Linear Models

Wu-Chyuan (Gary) Gau

Department of Statistics and Actuarial Science

University of Central Florida

June 17, 2009

Outline

- Exponential Family
- Components of GLM
- Estimation
- Inference
- Prediction
Exponential family

\[ f (y; \theta) = \exp \left[ a (y) b (\theta) + c (\theta) + d (y) \right], \]
where \(a, b, c, d\) are known functions.

- If \(a (y) = y\), the distribution is said to be in canonical (that is, standard) form.

- \(b (\theta)\) is sometimes called the natural parameter of the distribution.

- Other parameters, in addition to the parameter of interest \(\theta\) are regarded as nuisance parameters, treated as known.

- Poisson, Normal, and Binomial

Poisson Distribution

\[ f (y; \theta) = \frac{\theta^y e^{-\theta}}{y!}, \]
where \(y = 0, 1, 2, \cdots\)

- \(f (y; \theta) = \exp (y \log \theta - \theta - \log y! )\)

- \(a (y) = y\)

- \(b (\theta) = \log \theta\)
Normal Distribution

\[ f (y; \mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[ -\frac{1}{2\sigma^2} (y - \mu)^2 \right] , \]

where \( \mu \) is the parameter of interest and \( \sigma^2 \) is regarded as a nuisance parameter.

\[ f (y; \mu) = \exp \left[ -\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log (2\pi\sigma^2) \right] \]

- \( a (y) = y \)
- \( b (\mu) = \mu / \sigma^2 \)
- \( c (\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log (2\pi\sigma^2) \)
- \( d (y) = -\frac{y^2}{2\sigma^2} \)

Binomial Distribution

\[ f (y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \]

\[ = \frac{n!}{y! (n-y)!} \pi^y (1 - \pi)^{n-y} , \]

where \( y = 0, 1, 2, \cdots, n. \)

\[ f (y; \pi) = \exp \left[ y \log \pi - y \log (1 - \pi) + n \log (1 - \pi) + \log \left( \frac{n}{y} \right) \right] \]

- \( a (y) = y \)
- \( b (\pi) = \log \pi - \log (1 - \pi) = \log \left( \frac{\pi}{1-\pi} \right) \)
Properties

\[ E [a(Y)] = \frac{-c'(\theta)}{b'(\theta)} \]

\[ \text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \]

Log-likelihood Function

\[ l(\theta; y) = a(y)b(\theta) + c(\theta) + d(y) \]

\[ U(\theta; y) = \frac{dl(\theta; y)}{d\theta} = a(y)b'(\theta) + c'(\theta), \text{ called score statistics} \]

\[ U = a(Y)b'(\theta) + c'(\theta) \text{ regarded as a random variable (used for inference about parameters in GLMs)} \]

\[ E[U] = 0 \]

\[ \text{Var}[U] = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta) = \mathcal{S}, \text{ called the information} \]

\[ E[U'] = -\text{Var}(U) = -\mathcal{S} \]
Components of GLM

1. Random component: the probability distribution of the response variable $Y$

2. Systematic component: a linear combination of explanatory variables

3. Link function: an equation linking the expected value of $Y$ with a linear combination of explanatory variables

Random Component

1. Response variables $Y_1, \ldots, Y_N$ are assumed to share the same distribution form from the exponential family. That is

$$f (y_i; \theta_i) = \exp \left[ y_i b (\theta_i) + c (\theta_i) + d (y_i) \right].$$

2. The joint probability density function of $Y_1, \ldots, Y_N$ is

$$f (y_1, \ldots, y_N; \theta_1, \ldots, \theta_N) = \prod_{i=1}^{N} \exp \left[ y_i b (\theta_i) + c (\theta_i) + d (y_i) \right] = \exp \left[ \sum_{i=1}^{N} y_i b (\theta_i) + \sum_{i=1}^{N} c (\theta_i) + \sum_{i=1}^{N} d (y_i) \right]$$
Systematic Component

- A set of parameters $\beta$ and explanatory variables

$$X = \begin{bmatrix}
    x_1^T \\
    \vdots \\
    x_N^T
\end{bmatrix}
= \begin{bmatrix}
    x_{11} & \cdots & x_{1p} \\
    \vdots & \ddots & \vdots \\
    x_{N1} & \cdots & x_{Np}
\end{bmatrix}
$$

- $\beta^T = (\beta_1, \ldots, \beta_p)$, where $p < N$.

- A linear predictor

$$\eta_i = x_i^T \beta = \sum_{j=0}^{p} \beta_j x_{ij}.$$
Normal Linear Model

- $Y_i \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$.

- $g(\mu_i) = \mu_i = x_i^T\beta$, where the link function is the identity function.

- Usually, written as
  
  $$y = X\beta + e$$

  where $e^T = [e_1, \ldots, e_N]$ and $e_i$ i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots, N$.

- The linear component $\mu = X\beta$ represents the signal, and $e$ represents the noise, error, or random variation.

Logistic Regression

- $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_i)$.

- The joint probability of $Y_1, \ldots, Y_N$ is given by
  
  $$\prod_{i=1}^{N} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

  $$= \exp \left[ \sum_{i=1}^{N} y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^{N} \log(1 - \pi_i) \right]$$

- The link function is
  
  $$g(\pi_i) = \log \left( \frac{\pi_i}{1 - \pi_i} \right),$$

  the logit function.
Poisson Regression

- $Y_i \overset{\text{indep.}}{\sim} Possion(\lambda_i)$.

- The joint probability of $Y_1, \ldots, Y_N$ is given by
  \[
  \prod_{i=1}^{N} \frac{\lambda_i^{y_i}e^{-\lambda_i}}{y_i!} \\
  = \exp \left( \sum_{i=1}^{N} y_i \log \lambda_i - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \log y_i! \right)
  \]

- The link function is
  \[
  g(\lambda_i) = \log \lambda_i.
  \]

Estimation

- Consider independent random variables $Y_1, \ldots, Y_N$.

- $E[Y_i] = \mu_i$.

- $g(\mu_i) = x_i^T \beta$.

- For each $Y_i$, the log-likelihood function is
  \[
  l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i).
  \]

- $E(Y_i) = \mu_i = -c'(\theta_i) / b'(\theta_i)$

- $\text{Var}(Y_i) = \left[ b''(\theta)c'(\theta) - b''(\theta)b'(\theta) \right] / [b'(\theta)]^3$
• The log-likelihood function for all the $Y_i$’s is

\[
l = \sum_{i=1}^{N} l_i = \sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i).
\]

Maximum Likelihood Estimation

• To obtain the maximum likelihood estimator for the parameter $\beta_j$, we need

\[
\frac{\partial l}{\partial \beta_j} = U_j
\]

\[
= \sum_{i=1}^{N} \left[ \frac{\partial l_i}{\partial \beta_j} \right]
\]

\[
= \sum_{i=1}^{N} \left[ \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \right]
\]

\[
= \sum_{i=1}^{N} \left[ (y_i - \mu_i) x_{ij} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \right]
\]

using the chain rule.
• Since
\[ \frac{\partial l_i}{\partial \theta_i} = y_i b' (\theta_i) + c' (\theta_i) = b' (\theta_i) [y_i - \mu_i] \]
\[ \frac{\partial \mu_i}{\partial \theta_i} = -c'' (\theta_i) b' (\theta_i) + c' (\theta_i) b'' (\theta_i) \]
\[ = b' (\theta_i) \text{Var} (Y_i) \]
\[ \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}. \]

• The variance-covariance matrix of the $U'_j$s has terms
\[ \mathcal{S}_{jk} = E \left[ U_j U_k \right] \]
\[ = \sum_{i=1}^{N} x_{ij} x_{ik} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2, \]
which form the information matrix $\mathcal{S}$.

Newton-Raphson Algorithm

• The estimating equation is given by
\[ \mathcal{S}(m-1) b^{(m)} = \mathcal{S}(m-1) b^{(m-1)} + U^{(m-1)}, \quad (1) \]
where $b^{(m)}$ is the vector of estimates of the parameters $\beta_1, \ldots, \beta_p$ at the $m^{th}$ iteration, $\mathcal{S}(m-1)$ is the information matrix at the $(m - 1)^{th}$ iteration, $U^{(m-1)}$ is the vector of $U'_j$s at the $(m - 1)^{th}$ iteration.

• Note that $\mathcal{S}$ can be written as
\[ \mathcal{S} = X^T W X, \]
where $W$ is the $N \times N$ diagonal matrix with elements
\[ w_{ii} = \frac{1}{\text{Var} (Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2. \]
• Also, \( \mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)} \) is the vector with \( j^{th} \) element given by

\[
\frac{p}{N} \sum_{j=1}^{p} \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\text{Var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 b_k^{(m-1)} + \sum_{i=1}^{N} \left[ \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} x_{ij} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \right].
\]

• Thus, (2) can be written as

\[
\mathbf{X}^T \mathbf{Wz},
\]

where \( \mathbf{z} \) has elements

\[
z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + \left( y_i - \mu_i \right) \left( \frac{\partial \mu_i}{\partial \eta_i} \right),
\]

with \( \mu_i \) and \( \frac{\partial \mu_i}{\partial \eta_i} \) evaluated at \( \mathbf{b}^{(m-1)} \).

• Hence (1) can be written as

\[
\left( \mathbf{X}^T \mathbf{WX} \right)^{(m-1)} \mathbf{b}^{(m)} = \mathbf{X}^T \mathbf{Wz}^{(m-1)}.
\]

Simple Poisson Regression Example

• Data (Dobson and Barnett, 2008)

<table>
<thead>
<tr>
<th>( y_i )</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_i )</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• Assume that the responses \( Y_i \) are Poisson random variables.

• For illustrative purpose, let us model the relationship between \( Y_i \) and \( x_i \) by the straight line

\[
E(Y_i) = \mu_i = \beta_1 + \beta_2 x_i = x_i^T \beta,
\]

where

\[
\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \text{ and } x_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}
\]

for \( i = 1, \ldots, N \).
• That is we take the link function to be the identity function
\[ g(\mu_i) = \mu_i = x_i^T \beta = \eta_i. \]

• We have,
\[ \frac{\partial \mu_i}{\partial \eta_i} = 1. \]

• Thus,
\[ w_{ii} = \frac{1}{\text{Var}(Y_i)} = \frac{1}{\beta_1 + \beta_2 x_i} \]
and
\[ z_i = b_1 + b_2 x_i + (y_i - b_1 - b_2 x_i) = y_i. \]

• Also
\[ \Omega = X^T W X \]
\[ = \begin{bmatrix} \sum_{i=1}^{N} \frac{1}{b_1 + b_2 x_i} \\ \sum_{i=1}^{N} \frac{x_i}{b_1 + b_2 x_i} \\ \sum_{i=1}^{N} \frac{x_i^2}{b_1 + b_2 x_i} \end{bmatrix} \]

and
\[ X^T W z = \begin{bmatrix} \sum_{i=1}^{N} \frac{y_i}{x_i} \\ \sum_{i=1}^{N} \frac{x_i y_i}{x_i} \end{bmatrix}. \]

• The maximum likelihood estimates are obtained iteratively from the equations
\[ (X^T W X)^{(m-1)} b^{(m)} = X^T W z^{(m-1)}, \]
where the superscript \((m-1)\) denotes evaluation at \(b^{(m-1)}\).

• For this data, \(N = 9\),
\[ y = z = \begin{bmatrix} 2 \\ 3 \\ \vdots \\ 15 \end{bmatrix} \]
and

\[
X = \begin{bmatrix}
1 & -1 \\
1 & -1 \\
\vdots & \vdots \\
1 & 1
\end{bmatrix}.
\]

- Choose initial estimates \(b_1^{(1)} = 7\) and \(b_2^{(1)} = 5\).

- Therefore,

\[
(X^T W X)^{(1)} = \begin{bmatrix}
1.821429 & -0.75 \\
-0.75 & 1.25
\end{bmatrix},
\]

\[
X^T W z^{(1)} = \begin{bmatrix}
9.869048 \\
0.583333
\end{bmatrix},
\]

so

\[
b^{(2)} = \left( (X^T W X)^{(1)} \right)^{-1} X^T W z^{(1)} = \begin{bmatrix}
0.729167 & 0.4375 \\
0.4375 & 1.0625
\end{bmatrix} \begin{bmatrix}
9.869048 \\
0.583333
\end{bmatrix} = \begin{bmatrix}
7.4514 \\
4.9375
\end{bmatrix}.
\]

- The iterative process is continued until it converges.

<table>
<thead>
<tr>
<th>(m)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b_1^{(m)})</td>
<td>7</td>
<td>7.45139</td>
<td>7.45163</td>
<td>7.45163</td>
</tr>
<tr>
<td>(b_2^{(m)})</td>
<td>5</td>
<td>4.93750</td>
<td>4.93531</td>
<td>4.93530</td>
</tr>
</tbody>
</table>

- The maximum likelihood estimates are

\[
\hat{\beta} = \begin{bmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2
\end{bmatrix} = \begin{bmatrix}
7.45163 \\
4.93530
\end{bmatrix}.
\]

- The variance-covariance matrix for \(\hat{\beta}\) is the inverse of the information matrix \(\Im = X^T W X\). That is

\[
\Im^{-1} = \begin{bmatrix}
0.7817 & 0.4166 \\
0.4166 & 1.1863
\end{bmatrix}.
\]

- Thus the estimated standard error for \(\hat{\beta}_1\) is \(\sqrt{0.7817} = 0.8841\) and the estimated standard error for \(\hat{\beta}_2\) is \(\sqrt{1.1863} = 1.0892\).
So, for example, an approximate 95% confidence interval for the slope $\beta_2$ is

$$4.93530 \pm 1.96 \times 1.0892 \text{ or } (2.80, 7.07).$$

R Code (Poisson Regression)

```
> res.p = glm(y~x, family=poisson(link="identity"))
> summary(res.p)
```
Logistic Regression

- Beetle mortality data (Bliss, 1935)

<table>
<thead>
<tr>
<th>Dose, $x_i$ ($\log_{10} CS_2 mg l^{-1}$)</th>
<th>Number of beetle, $n_i$</th>
<th>Number killed, $y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

- Consider $N$ random variables $Y_1, \ldots, Y_N$, where $Y_i \sim Bin(n_i, \pi_i)$.

- The log-likelihood function is

$$l(\pi_1, \ldots, \pi_N; y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[ y_i \log \left( \frac{\pi_i}{1-\pi_i} \right) + n_i \log \left( 1-\pi_i \right) + \log \left( \frac{n_i}{y_i} \right) \right].$$

- Fitting the logistic model, so

$$\log \left( \frac{\pi_i}{1-\pi_i} \right) = \beta_1 + \beta_2 x_i,$$

so

$$\pi_i = \frac{\exp (\beta_1 + \beta_2 x_i)}{1 + \exp (\beta_1 + \beta_2 x_i)}$$

and

$$\log (1 - \pi_i) = - \log [1 + \exp (\beta_1 + \beta_2 x_i)].$$

- Therefore, the log-likelihood function is

$$l = \sum_{i=1}^{N} \left[ y_i (\beta_1 + \beta_2 x_i) - n_i \log [1 + \exp (\beta_1 + \beta_2 x_i)] + \log \left( \frac{n_i}{y_i} \right) \right].$$
• The link function is
\[ g(\pi_i) = \log \left( \frac{\pi_i}{1 - \pi_i} \right). \]

• The scores with respect to \( \beta_1 \) and \( \beta_2 \) are
\[ U_1 = \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{N} (y_i - n_i\pi_i) \]
\[ U_2 = \frac{\partial l}{\partial \beta_2} = \sum_{i=1}^{N} x_i (y_i - n_i\pi_i). \]

• The information matrix is
\[ \mathcal{I} = \begin{bmatrix} \sum_{i=1}^{N} n_i\pi_i (1 - \pi_i) & \sum_{i=1}^{N} n_i x_i \pi_i (1 - \pi_i) \\ \sum_{i=1}^{N} n_i x_i \pi_i (1 - \pi_i) & \sum_{i=1}^{N} n_i x_i^2 \pi_i (1 - \pi_i) \end{bmatrix}. \]

• Maximum likelihood estimates are obtained by solving the iterative equation
\[ \mathcal{I}^{(m-1)}b^{(m)} = \mathcal{I}^{(m-1)}b^{(m-1)} + U^{(m-1)}. \]

• The iterative process is continued until it converges.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_1^{(m)} )</td>
<td>0</td>
<td>-37.856</td>
<td>-53.853</td>
<td>-60.717</td>
</tr>
<tr>
<td>( b_2^{(m)} )</td>
<td>0</td>
<td>21.337</td>
<td>30.384</td>
<td>34.270</td>
</tr>
</tbody>
</table>
R Code

Data entry and manipulation

```r
> y = c(6, 13, 18, 28, 52, 53, 61, 60)
> n = c(59, 60, 62, 56, 63, 59, 62, 60)
> x = c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
> n_y = n - y
> beetle.mat = cbind(y, n_y)
```

Logistic Regression

```r
> res.glm = glm(beetle.mat ~ x,
               family = binomial(link = "logit"))
```

Estimated Proportion of Deaths

```r
> fit_p = c(fitted.values(res.glm))
> fit_y = n * fit_p
```
Inference

- Confidence Intervals
- Hypothesis Tests

Statistic

- Wald Statistic:
  \[(b - \beta)^T \Sigma (b) (b - \beta) \sim \chi^2 (p),\]
  where \(p\) is the number of parameters in the model.

  - e.g., one-parameter case, the more commonly used form is
    \[b \sim N (\beta, \Sigma^{-1})\]

- Log-likelihood ratio statistic (or deviance): measure how well the models fit the data (the goodness of fit).
  \[D = 2 \left[ l (b_{\text{max}}; y) - l (b; y) \right] \sim \chi^2 (m - p, v),\]

  - e.g., if Normally distributed, \(D\) has a chi-squared distribution exactly.
Hypothesis Test

1. Specify a model $M_0$ corresponding to $H_0$. Specify a more general model $M_1$ (with $M_0$ as a special case of $M_1$).

2. Fit $M_0$ and calculate the goodness of fit statistic $G_0$. Fit $M_1$ and calculate the goodness of fit statistic $G_1$.

3. Calculate the improvement in fit, usually $G_1 - G_0$ but $G_1 / G_0$ is another possibility.

4. Use the sampling distribution of $G_1 - G_0$ (or some related statistics) to test $H_0 : G_1 = G_0$ against $H_0 : G_1 \neq G_0$.

Prediction

- Consider $(y_i, x_{i1}, \ldots, x_{ip})$ in the logistic regression.

- The predicted probability can be computed from

$$P \{ Y_i = 1 | x_i \} = \pi (x_i) \frac{\exp (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}{1 + \exp (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}.$$  

- To obtain the derived dichotomous variable, we compare each estimated probability to a cutoff point $c$.  

- The predicted binary outcome is then given by

$$\hat{y}_i = 1 \{ \hat{\pi}_i \geq c \}.$$
Prediction Accuracy

- Classification Table: For a given threshold $c \in [0, 1]$, we can form the following $2 \times 2$ classification table.

<table>
<thead>
<tr>
<th>Observed</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$n_{00}$</td>
<td>$n_{01}$</td>
<td>$n_0$.</td>
</tr>
<tr>
<td>1</td>
<td>$n_{10}$</td>
<td>$n_{11}$</td>
<td>$n_1$.</td>
</tr>
<tr>
<td>Total</td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$n$.</td>
</tr>
</tbody>
</table>

- Sensitivity=$n_{11}/n_1$, the proportion of correctly-predicted events.

- Specificity=$n_{00}/n_0$, the proportion of correctly-predicted non-events.

- Prediction Accuracy=$(n_{00} + n_{11})/n$, the proportion of correctly classified subjects.

Selection of Cutoff Point $c$

- By default, the cutoff point $c = 0.5$.

- Or, find the optimal cutoff point

![Graphs showing sensitivity and specificity](image-url)
Summary

- Generalized Linear Models (GLM; McCullagh and Nelder, 1983) extend the ordinary linear regression model to encompass nonnormal response while, at the same time, enjoying nearly all its merits.

- Within the maximum likelihood framework, GLMs provide a unified approach for commonly used linear models with continuous, binary/categorical, or count response.

- GLMs are flexible and easy to implement.

- Actuaries are capable.