Predictive Modeling Symposium  
October 8-9, 2009  

Session 1a: An Introduction to Two-Part Models and Longitudinal Models for Use in Modeling Health Care Utilization  

Edward W. Frees  
Marjorie Rosenberg, Ph.D., FSA, MAAA
An Introduction to Two-Part Models and Longitudinal Models for Use in Modeling Health Care Utilization

Edward (Jed) W. Frees and Marjorie A. Rosenberg

Department of Actuarial Science, Risk Management and Insurance
School of Business
University of Wisconsin – Madison

October 8, 2009

Outline I

1 Introduction
2 Introduction to Two-Part Models
   ● Concept
   ● Refresher on Regression
3 Description of Data
   ● Overview
   ● Available Data
4 Analysis Approach
   ● Define Basic Inputs
   ● Modeling of Data
   ● Covariates
5 Descriptive Statistics and Modeling

Outline II

1 Descriptive Statistics
2 Introduction to Tobit Model
3 Reminder of Two-Part Model
4 SAS Modeling
5 Another Approach: Aggregate Loss Models
6 Brief Introduction to Longitudinal Models
   ● Longitudinal Data
   ● Motivating Example
   ● Descriptive Statistics
   ● Longitudinal Results
7 Seminar Conclusion

Handouts and Support Material

- This Presentation – TwoPartOct2009Print.pdf
- Tables of Results – Chap16Tables.pdf
- NAAJ case study paper – NAAJ0703-3.pdf
- Raw SAS code – Chapter16Tables.sas and MEPSExtraction.sas
- Two Data Sets – HealthExpend.SAS7BDAT, HealthExpendEvent.SAS7BDAT
Background of Presenters

- Margie Rosenberg: PhD, FSA with joint appointment in Business and Biostatistics and Medical Informatics. Research interests in modeling healthcare utilization.

Introduction of Participants

Objective of Workshop

To introduce participants to the basics of two different advanced statistical methods and allow interactive use in SAS during the workshop

- Two-part modeling
- Longitudinal modeling

Considerations

- Healthcare data often feature a large proportion of zeros, where zero values can represent:
  - Individual's lack of healthcare utilization,
  - No expenditure, or
  - Non-participation in a program.

- How to model zero expenditures?
  - Ignore that they exist.
  - 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.
- $(\text{Claim recorded})_i = r_i \times y_i$

Two-part models include covariates in each part.

**Medical Expenditure Panel Survey (MEPS)**

- Conducted by Agency for Health Care Research and Quality (AHRQ) and National Center for Health Statistics (NCHS)
- Uses National Health Interview Survey (NHIS) as sampling frame
- 2 year panel
- 5 computer-assisted personal interviews (CAPI) to collect 2 full years of data
- Overlapping data collection
- 9 panels between 1996-2004
MEPS Family of Surveys

- Household component (HC)-core survey
- Medical provider component (MPC)
- HC and MPC form the basis for national estimates of healthcare use and expenditures
- Insurance component (IC)
- Nursing home component (NHC)-1996 only
- Focus on HC and MPC

MEPS - Household Component

- Collects population characteristics files
  - Person and family level
  - Full-year or point-in-time
- Available data
  - Demographic
  - Income
  - Health status
  - Disability days
  - Access to care
  - Employment
  - Health insurance
  - Utilization, expenditures and sources of payments variables

MEPS - Household Component

Medical Data

- Prescription medicine
- Dental visits
- Other medical expenses
- Hospital inpatient stays
- Emergency room visits
- Outpatient department visits
- Office-based medical provider visits
- Home health

The Modeling Process, Klugman et al. (p. 4)
### Task

**Goal**

Predict the counts and amounts of expenditures for inpatient utilization.

**Data**

MEPS includes inpatient events include pharmaceutical, procedures, medical care provided, room and board, and emergency room visits that result in inpatient stay.

**Method**

Ordinary Regression, Tobit and Two-Part Models

### Variables Considered

- Education
- Demographic
- Geographic
- Health Status
- Economic Factors

### Education Covariate

Education has an ambiguous impact on the demand for health care services.

- **Theory 1**: More educated persons more aware of health risks, thus more active in maintaining their health.
- **Theory 2**: Less educated persons greater exposure to health risks and develop a greater immunity for certain types of risks.

In MEPS, education proxied by number of degrees received and categorized into three different levels: lower than high school, high school, and college or above education.

### Demographic, Geographic, Health Status Covariates

- Health deteriorates with age.
- Sex and ethnicity proxies for inherited health and social habits in maintaining health.
- Geographical region proxy for accessibility of health care services and the overall economic or regional impact on residents’ health care behavior.
- Self-rated physical health, mental health and any functional or activity related limitations during the sample period are used as proxies for health status.
Economic Covariates

Economic factors include income and insurance coverage.

Income in MEPS: Measured relative to the poverty line.

Insurance Coverage: Important variable in explaining health care utilization.

(i) Reduces out-of-pocket by insureds and thus induces moral hazard.

(ii) Rand Health Insurance Experiment empirically suggested that cost sharing effects affect primarily the number of medical contacts, rather than the intensity of each contact.

(iii) We use a binary variable:

\[
Y_i = \begin{cases} 
1 & \text{If at least one month of insurance per year} \\
0 & \text{Otherwise} 
\end{cases}
\]

Tobit Model

Why not ignore zeros?

With the normality assumption, standard calculations show that

\[
E[y_i] = d_i + \Phi \left( \frac{x_i'\beta - d_i}{\sigma} \right) (x_i'\beta - d_i + \sigma \lambda_i), \quad \text{where}
\]

\[
\lambda_i = \frac{\phi \left( (x_i'\beta - d_i)/\sigma \right)}{\Phi \left( (x_i'\beta - d_i)/\sigma \right)}.
\]

Thus, usual regression coefficients are biased. If \( \frac{x_i'\beta - d_i}{\sigma} \) is large, then the bias is small.
Tobit Model

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 \( \{ i : y_i = d_i \} = \text{Sum of censored observations} \) and \( \{ i : y_i > d_i \} = \text{Sum over non-censored observations.} \)

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.

Ordinary Regression and Tobit Model Fit

- Descriptive Statistics
- Ordinary Regression, Tobit and Two-Part Models

<table>
<thead>
<tr>
<th>Effect</th>
<th>OLS Parameter Estimate</th>
<th>t-value</th>
<th>Tobit Parameter Estimate</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.123</td>
<td>-0.525</td>
<td>-0.123</td>
<td>-0.525</td>
</tr>
<tr>
<td>AGE</td>
<td>0.001</td>
<td>0.091</td>
<td>-0.006</td>
<td>-0.118</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.379</td>
<td>3.711</td>
<td>5.727</td>
<td>4.107</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.115</td>
<td>-0.459</td>
<td>-1.732</td>
<td>-0.480</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.054</td>
<td>0.365</td>
<td>0.025</td>
<td>0.015</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.359</td>
<td>0.726</td>
<td>3.745</td>
<td>0.723</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>1.702</td>
<td>3.828</td>
<td>1.849</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.265</td>
<td>1.693</td>
<td>3.459</td>
<td>1.790</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.146</td>
<td>1.133</td>
<td>1.805</td>
<td>1.056</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>-0.014</td>
<td>-0.089</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>-0.027</td>
<td>-0.209</td>
<td>-0.030</td>
<td>-0.019</td>
</tr>
<tr>
<td>POOR</td>
<td>0.207</td>
<td>1.513</td>
<td>13.352</td>
<td>4.838</td>
</tr>
<tr>
<td>FAIR</td>
<td>-0.001</td>
<td>-0.004</td>
<td>1.354</td>
<td>0.528</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.189</td>
<td>1.546</td>
<td>2.740</td>
<td>1.480</td>
</tr>
<tr>
<td>VGOOD</td>
<td>0.084</td>
<td>0.622</td>
<td>1.506</td>
<td>0.615</td>
</tr>
<tr>
<td>NHPOR</td>
<td>0.000</td>
<td>-0.001</td>
<td>-0.082</td>
<td>-0.211</td>
</tr>
<tr>
<td>ANY/LMT</td>
<td>0.415</td>
<td>3.103</td>
<td>4.695</td>
<td>3.000</td>
</tr>
<tr>
<td>HINCOME</td>
<td>0.482</td>
<td>3.718</td>
<td>8.575</td>
<td>3.955</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.393</td>
<td>-1.669</td>
<td>-4.265</td>
<td>-2.241</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.175</td>
<td>-0.976</td>
<td>-3.414</td>
<td>-1.619</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.116</td>
<td>-0.478</td>
<td>-2.274</td>
<td>-0.799</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.594</td>
<td>4.486</td>
<td>8.534</td>
<td>4.130</td>
</tr>
<tr>
<td>Scale ( \sigma )</td>
<td>4.999</td>
<td>14.738</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Two-Part Model Fit

<table>
<thead>
<tr>
<th>Effect</th>
<th>Frequency</th>
<th>Severity</th>
<th>Frequency</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.932</td>
<td>-0.678</td>
<td>-2.817</td>
<td>-1.342</td>
</tr>
<tr>
<td>AGE</td>
<td>0.002</td>
<td>0.012</td>
<td>0.012</td>
<td>1.368</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.395</td>
<td>-0.104</td>
<td>0.395</td>
<td>-0.102</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.108</td>
<td>-0.429</td>
<td>-0.108</td>
<td>-0.427</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.008</td>
<td>0.088</td>
<td>0.009</td>
<td>0.073</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.284</td>
<td>-0.639</td>
<td>0.285</td>
<td>-0.780</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>-0.649</td>
<td>0.281</td>
<td>-0.778</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.239</td>
<td>0.016</td>
<td>0.237</td>
<td>1.754</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.132</td>
<td>-0.078</td>
<td>0.130</td>
<td>-0.065</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.048</td>
<td>0.056</td>
<td>0.049</td>
<td>0.062</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>0.002</td>
<td>-0.415</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>POOR</td>
<td>0.955</td>
<td>0.597</td>
<td>0.939</td>
<td>4.805</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.087</td>
<td>-0.211</td>
<td>0.079</td>
<td>0.450</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.184</td>
<td>1.422</td>
<td>0.182</td>
<td>1.412</td>
</tr>
<tr>
<td>VG00</td>
<td>0.095</td>
<td>0.373</td>
<td>0.094</td>
<td>0.728</td>
</tr>
<tr>
<td>MNMPPOOR</td>
<td>-0.027</td>
<td>-0.164</td>
<td>-0.176</td>
<td>-0.579</td>
</tr>
<tr>
<td>ANYUMT</td>
<td>0.318</td>
<td>2.941</td>
<td>0.311</td>
<td>3.022</td>
</tr>
<tr>
<td>TRACOMC</td>
<td>0.682</td>
<td>3.131</td>
<td>0.679</td>
<td>3.494</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.314</td>
<td>-2.316</td>
<td>-0.314</td>
<td>-2.345</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.241</td>
<td>-1.626</td>
<td>-0.241</td>
<td>-1.633</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.145</td>
<td>-0.716</td>
<td>-0.146</td>
<td>-0.721</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.580</td>
<td>4.154</td>
<td>0.579</td>
<td>4.147</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>1.999</td>
<td></td>
<td>1.999</td>
<td></td>
</tr>
</tbody>
</table>

### Aggregate Loss Model Fit Statistics

<table>
<thead>
<tr>
<th>Effect</th>
<th>Frequency</th>
<th>Severity</th>
<th>Frequency</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.214</td>
<td>-0.619</td>
<td>9.742</td>
<td>15.514</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.005</td>
<td>-0.756</td>
<td>-0.006</td>
<td>-0.747</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.671</td>
<td>3.351</td>
<td>-0.385</td>
<td>-1.952</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.153</td>
<td>-0.306</td>
<td>-0.340</td>
<td>-0.588</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.144</td>
<td>0.639</td>
<td>0.146</td>
<td>0.686</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.445</td>
<td>0.624</td>
<td>-0.331</td>
<td>-0.465</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.492</td>
<td>1.683</td>
<td>-0.547</td>
<td>-1.792</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.619</td>
<td>2.314</td>
<td>0.303</td>
<td>-0.070</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.391</td>
<td>1.603</td>
<td>0.108</td>
<td>0.424</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.007</td>
<td>0.089</td>
<td>0.169</td>
<td>2.394</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>-0.085</td>
<td>-0.399</td>
<td>-0.722</td>
<td>-3.396</td>
</tr>
<tr>
<td>POOR</td>
<td>1.927</td>
<td>5.211</td>
<td>0.064</td>
<td>1.968</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.226</td>
<td>0.827</td>
<td>-0.186</td>
<td>-0.486</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.385</td>
<td>1.483</td>
<td>0.232</td>
<td>0.802</td>
</tr>
<tr>
<td>VG00</td>
<td>0.348</td>
<td>1.349</td>
<td>0.429</td>
<td>1.011</td>
</tr>
<tr>
<td>MNMPPOOR</td>
<td>-0.177</td>
<td>-0.583</td>
<td>-0.221</td>
<td>-0.870</td>
</tr>
<tr>
<td>ANYUMT</td>
<td>0.774</td>
<td>3.499</td>
<td>0.579</td>
<td>2.720</td>
</tr>
<tr>
<td>TRACOMC</td>
<td>-0.662</td>
<td>-2.197</td>
<td>-0.725</td>
<td>-2.371</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.482</td>
<td>-1.631</td>
<td>-0.720</td>
<td>-2.768</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.460</td>
<td>-1.611</td>
<td>-0.631</td>
<td>-2.341</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.465</td>
<td>-1.131</td>
<td>-0.056</td>
<td>-0.135</td>
</tr>
<tr>
<td>INSURE</td>
<td>1.312</td>
<td>4.207</td>
<td>1.500</td>
<td>4.551</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>2.177</td>
<td></td>
<td>1.314</td>
<td></td>
</tr>
</tbody>
</table>

### Definition of Aggregate Loss Model

- First part of two-part model re-defined to include multiple hospitalizations.
- Instead of binary variable for utilization, use $N_i$ as random variable with support of $0, 1, \ldots$.
- Choose Poisson or Negative Binomial random variable, or similar discrete random variable.
- Second part of two-part model as before, i.e. conditional on $N_i > 0$.
- Typical models include ordinary regression (using log transform), gamma regression and mixed models.
- For the mixed models, use subject-specific intercept to account for the heterogeneity among subjects.

---

### Longitudinal Data

- Longitudinal/panel data – data with “double subscripts.”
- There are $n$ independent subjects. We observe the $i$th subject over $t = 1, ..., T_i$ time periods, for each of $i=1, ..., n$ subjects.
- The response variable for the $i$th subject during the $tt$th time period is $Y_{it}$.
- The $k$ explanatory variables are $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itk})'$, a vector of dimension $k \times 1$.
Wisconsin Nursing Homes


- 398 nursing facilities over 7 years, 4,076 facility-year observations.

- Dependent variable is total patient years
  - Data varies both in cross-section and over time;
  - 1989 – 1999 in-sample data to develop model;

### Descriptive Statistics of Continuous Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPY</td>
<td>Total Patient Years (median 98.7)</td>
</tr>
</tbody>
</table>

### Continuous Covariates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEAR</td>
<td>Cost report year minus 1988 (1 to 13)</td>
</tr>
<tr>
<td>NumBed</td>
<td>Number of beds (median 95)</td>
</tr>
<tr>
<td>SqrFoot</td>
<td>Nursing home net square footage (in thousands of square feet, median 37.27)</td>
</tr>
</tbody>
</table>

### Descriptive Statistics of Categorical Covariates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>POPID</td>
<td>Nursing home identification number</td>
</tr>
<tr>
<td>SelfFundIns</td>
<td>Self Funding of Insurance</td>
</tr>
<tr>
<td>Yes</td>
<td>50.4</td>
</tr>
<tr>
<td>No</td>
<td>49.6</td>
</tr>
<tr>
<td>MCert</td>
<td>Medicare-certified</td>
</tr>
<tr>
<td>Yes</td>
<td>70.2</td>
</tr>
<tr>
<td>No</td>
<td>29.8</td>
</tr>
<tr>
<td>Organizational</td>
<td>Pro (nursing home is for profit)</td>
</tr>
<tr>
<td>Yes</td>
<td>52.4</td>
</tr>
<tr>
<td>No</td>
<td>47.6</td>
</tr>
<tr>
<td>Structure</td>
<td>TaxExempt (nursing home is tax exempt)</td>
</tr>
<tr>
<td>Yes</td>
<td>36.8</td>
</tr>
<tr>
<td>No</td>
<td>63.2</td>
</tr>
<tr>
<td>Location</td>
<td>Govt (nursing home is a governmental unit)</td>
</tr>
<tr>
<td>Yes</td>
<td>10.8</td>
</tr>
<tr>
<td>No</td>
<td>89.2</td>
</tr>
<tr>
<td>MSA</td>
<td>Metropolitan Statistical Area code (1 to 13, 0 rural)</td>
</tr>
<tr>
<td>Urban</td>
<td>53.8</td>
</tr>
<tr>
<td>Rural</td>
<td>46.2</td>
</tr>
<tr>
<td>Location</td>
<td>Urban</td>
</tr>
<tr>
<td>Yes</td>
<td>53.8</td>
</tr>
<tr>
<td>No</td>
<td>46.2</td>
</tr>
</tbody>
</table>

### Histogram of TPY and logarithmic TPY

Histograms showing the distribution of total patient years (TPY) and the logarithmic transformation of TPY.
### Summary Statistics of TPY by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Facilities</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>376</td>
<td>102.0</td>
<td>88.3</td>
<td>67.0</td>
<td>15.1</td>
<td>645.0</td>
<td>65.7</td>
</tr>
<tr>
<td>1990</td>
<td>376</td>
<td>103.4</td>
<td>88.5</td>
<td>66.9</td>
<td>14.2</td>
<td>650.6</td>
<td>64.7</td>
</tr>
<tr>
<td>1991</td>
<td>368</td>
<td>104.1</td>
<td>90.1</td>
<td>66.8</td>
<td>13.2</td>
<td>655.1</td>
<td>64.2</td>
</tr>
<tr>
<td>1992</td>
<td>372</td>
<td>105.2</td>
<td>89.3</td>
<td>67.5</td>
<td>16.0</td>
<td>653.0</td>
<td>64.2</td>
</tr>
<tr>
<td>1993</td>
<td>367</td>
<td>105.5</td>
<td>89.7</td>
<td>68.0</td>
<td>16.5</td>
<td>657.3</td>
<td>64.5</td>
</tr>
<tr>
<td>1994</td>
<td>370</td>
<td>104.4</td>
<td>89.4</td>
<td>67.2</td>
<td>14.7</td>
<td>658.6</td>
<td>64.4</td>
</tr>
<tr>
<td>1995</td>
<td>370</td>
<td>104.6</td>
<td>90.4</td>
<td>67.5</td>
<td>16.6</td>
<td>669.3</td>
<td>64.5</td>
</tr>
<tr>
<td>1996</td>
<td>372</td>
<td>101.8</td>
<td>87.8</td>
<td>65.3</td>
<td>15.7</td>
<td>657.3</td>
<td>64.1</td>
</tr>
<tr>
<td>1997</td>
<td>370</td>
<td>99.7</td>
<td>85.3</td>
<td>62.3</td>
<td>15.7</td>
<td>614.4</td>
<td>62.5</td>
</tr>
<tr>
<td>1998</td>
<td>365</td>
<td>97.2</td>
<td>85.1</td>
<td>59.5</td>
<td>11.9</td>
<td>551.5</td>
<td>61.2</td>
</tr>
<tr>
<td>1999</td>
<td>370</td>
<td>92.4</td>
<td>82.7</td>
<td>52.3</td>
<td>11.6</td>
<td>439.6</td>
<td>56.6</td>
</tr>
<tr>
<td>2000</td>
<td>362</td>
<td>88.8</td>
<td>80.5</td>
<td>46.1</td>
<td>11.6</td>
<td>314.7</td>
<td>51.9</td>
</tr>
<tr>
<td>2001</td>
<td>355</td>
<td>89.7</td>
<td>81.1</td>
<td>49.0</td>
<td>12.3</td>
<td>440.7</td>
<td>54.7</td>
</tr>
<tr>
<td>1989-2001</td>
<td>4793</td>
<td>100.0</td>
<td>86.7</td>
<td>62.7</td>
<td>11.6</td>
<td>669.3</td>
<td>62.7</td>
</tr>
</tbody>
</table>

### Observations from Validation

- Ordinary Regression model performed poorly compared to the other candidate models.
- Both Fixed Effects and Mixed Effects models performed better than the most recent observation, with the slight edge to Mixed Effects.
- Transforming TPY better than not transforming TPY.

### Overall Conclusions

- Actuaries have learned no-covariate form two-part model on exam syllabus.
- We re-introduced ordinary regression and put into context of two-part model for cross-sectional data.
- We showed example of power of longitudinal model that takes into correlations of data over time.
Goals
- To educate actuaries to be conversant in statistics and in advanced statistical techniques.
- To enable awareness of how the process of completing their statistical work could evolve.
- To effectively communicate the use of statistical techniques and the results of their application.

Quantification of Uncertainty: A Proposal of the Development of Executive Education Courses to Advance the Statistical Thinking of Actuaries

- An Overview of Statistical Methods for Actuaries
  - One day course geared towards Management-level
  - Refresher course on statistics grounded in examples and applications in the insurance industry
- A Case Study Approach for Advanced Statistical Methods for Actuaries
  - Two day advanced hands-on course
  - Interactive case study approach using computers
PREDICTIVE MODELING WITH LONGITUDINAL DATA: A CASE STUDY OF WISCONSIN NURSING HOMES

Marjorie A. Rosenberg, * Edward W. Frees, † Jiafeng Sun,‡
Paul H. Johnson, Jr.,/H14067 and James M. Robinson/H1606

ABSTRACT

The recent development and availability of sophisticated computer software has facilitated the use of predictive modeling by actuaries and other financial analysts. Predictive modeling has been used for several applications in both the health and property and casualty sectors. Often these applications employ extensions of industry-specific techniques and do not make full use of information contained in the data. In contrast, we employ fundamental statistical methods for predictive modeling that can be used in a variety of disciplines. As demonstrated in this article, this methodology permits a disciplined approach to model building, including model development and validation phases. This article is intended as a tutorial for the analyst interested in using predictive modeling by making the process more transparent.

This article illustrates the predictive modeling process using State of Wisconsin nursing home cost reports. We examine utilization of approximately 400 nursing homes from 1989 to 2001. Because the data vary both in the cross section and over time, we employ longitudinal models. This article demonstrates many of the common difficulties that analysts face in analyzing longitudinal health care data, as well as techniques for addressing these difficulties. We find that longitudinal methods, which use historical trend information, significantly outperform regression models that do not take advantage of historical trends.

1. INTRODUCTION

The recent development and availability of sophisticated computer software has facilitated the use of predictive modeling by actuaries and other financial analysts. Predictive modeling refers to a statistical process of analyzing data related to some problem of interest. This process can be described as (1) defining the problem to be studied, (2) collecting sufficient knowledge about the problem and obtaining appropriate data, (3) examining trends in the data to aid in developing candidate models (sometimes referred to as data mining), (4) estimating the candidate models via reasonable methods, and (5) using diagnostic analyses and selection criteria to decide which of the candidate models is best for

* Marjorie A. Rosenberg, FSA, MAAA, Ph.D. is an Associate Professor with a joint appointment in the Department of Actuarial Science, Risk Management and Insurance and the Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53706, mrosenberg@bus.wisc.edu.
† Edward W. Frees, FSA, MAAA, Ph.D. is a Professor in the Department of Actuarial Science, Risk Management and Insurance, University of Wisconsin–Madison, Madison, WI 53706, jfrees@bus.wisc.edu.
‡ Jiafeng Sun is a Ph.D. candidate in the Department of Actuarial Science, Risk Management and Insurance, University of Wisconsin–Madison, Madison, WI 53706, jiafengsun@wisc.edu.
/H14067 Paul H. Johnson, Jr. is a Ph.D. candidate in the Department of Actuarial Science, Risk Management and Insurance, School Of Business, University of Wisconsin–Madison, Madison, WI 53706, pauljohnson@wisc.edu.
/H1606 Jim Robinson, FSA, MAAA, Ph.D., is a Senior Scientist and Deputy Director of the Center for Health Systems Research and Analysis, University of Wisconsin–Madison, Madison, WI 53706-2397, jim@chsra.wisc.edu.
analyzing the problem (Rosenberg and Johnson 2007). This article is intended as a tutorial for the analyst interested in using longitudinal data for predictive modeling and focuses on the development, estimation, and validation of predictive models.

Predictive modeling has been used extensively in the health care sector; see, for example, Cumming and Cameron (2002) for an evaluation of several commercial health-care-related predictive modeling software packages. Many health-sector applications require that predictive models account for variables related to the health status of individuals or groups by including covariates like age, gender, race/ethnicity, diagnosis codes, cultural and socioeconomic attributes, and health assessment measures such as quality of life and self-reported health. Accounting for health-status-related covariates is known as risk adjustment (Iezzoni 1997). There are four major health-sector applications of predictive modeling. One application is provider profiling, where provider performances are ordered based on the quality of treatment, number of tests, and disease severity of their case mix (Christiansen and Morris 1997; Delong et al. 1997; Hu and Lesneski 2004). A second application is provider reimbursement, where providers who treat Medicare or Medicaid insured patients receive payment as determined by a statistical model (Ash and Byrne-Logan 1998; Ash et al. 2000; Pope et al. 2000; Kronick et al. 2000). A third application is the identification of individual or groups of patients that are likely to be high-cost users of medical services in future periods, for the purpose of targeting them with interventions to reduce future costs (Cousins, Shickle, and Bander 2002; Passwater and Seiler 2004; Dove, Duncan, and Robb 2003; Meenan et al. 1999; Zhao et al. 2003). Finally, predictive modeling is used to supplement the underwriting and pricing of small group health insurance (Cumming et al. 2002; Ellis et al. 2003; Hu and Lesneski 2004).

Predictive modeling also has been utilized in the property-casualty sector. Derrig (2002) provided a general overview of insurance fraud detection and deterrence strategies, and other analysts have developed predictive models for detecting and classifying types of insurance fraud (Brockett et al. 2002; Tennyson and Slasas-Forn 2002; Viaene et al. 2002). Predictive modeling also has been used to relate credit scores to personal automobile or homeowners' profitability (Monaghan 2000; Wu and Guszeza 2003). The American Academy of Actuaries summarized four studies that discussed the use of credit history for personal lines of insurance (American Academy of Actuaries Risk Classification Subcommittee of the Property/Casualty Products and Committee 2002). Medical malpractice claims and litigation have been analyzed with predictive models (Cooil 1991; Weycker and Jensen 2000). Claims reserving naturally lends itself to predictive modeling (Guszeza and Lommele 2006). Finally, predictive modeling can assist in predicting claimant behavior in workers’ compensation (Biddle and Roberts 2003; Speights, Brodsky, and Chudova 1999).

This article illustrates the predictive modeling process using data from State of Wisconsin nursing home cost reports. We examine utilization of approximately 400 nursing homes from 1989 to 2001 and are interested in forecasting total patient years (number of total patient days in the cost reporting period divided by number of facility operating days in the cost reporting period) by individual nursing home. The Wisconsin nursing home data vary both in the cross-section and over time; therefore, we use longitudinal (panel data) models. Covariates are included as with ordinary regression, yet differences between subjects and changes over time can be incorporated. See Baltagi (2005), Diggle et al. (2002), or Frees (2004) for a general discussion of longitudinal data modeling. This article demonstrates how poorly cross-sectional methods perform because they do not take advantage of historical trends when predicting future outcomes, as compared to longitudinal methods that use historical trend information. This article also demonstrates many of the common difficulties that analysts face in analyzing longitudinal health care data, as well as techniques for addressing these difficulties. The longitudinal data approach used in this article is only one of several predictive modeling approaches; other approaches are discussed in the Summary section.

The data are used to illustrate the modeling techniques. The empirical results obtained provide projections about future nursing home utilization. However, any general conclusions regarding nursing homes/custodial care would consider the economic, financial, and legal environments in which the
nursing homes operate and would incorporate related variables (such as the demand and supply of beds, the degree of competition within the market, Wisconsin poverty levels, and Wisconsin Medicaid legislation).

The remainder of the article is organized as follows. Section 2 describes the nursing home data. Section 3 provides summary statistics and develops and estimates candidate longitudinal models. Section 4 validates these candidate models by predicting future outcomes. Section 5 provides a summary and concluding remarks. An appendix provides details of the model selection and sample statistical code.

2. CASE STUDY ON NURSING HOME DATA

The State of Wisconsin Medicaid program funds nursing home care for individuals qualifying on the basis of need and financial status. Nursing home care providers are categorized into three types: nursing facilities providing skilled care to disabled adults and frail elderly, facilities serving people with developmental disabilities (FDDs), and facilities serving both populations in separate units. Most, but not all, nursing homes in Wisconsin are certified to provide Medicaid-funded care. Those that do not accept Medicaid are generally paid directly by the resident or the resident’s insurer. Most, but not all, nursing facilities are certified to provide Medicare-funded care. Medicare provides postacute care for 100 days following a related hospitalization. Medicare does not fund care provided by FDDs.

Nursing homes are owned and operated by a variety of entities, including the state, counties, municipalities, for-profit businesses, and tax-exempt organizations. Private firms often own several nursing homes. Periodically facilities may change ownership and, less frequently, ownership type. Some nursing homes opt not to purchase private insurance coverage for their employees. Instead, these facilities directly provide insurance and pension benefits to their employees; this is referred to as “self-funding of insurance.”

As part of the conditions for participation, Medicaid-certified nursing homes must file an annual cost report to the Wisconsin Department of Health and Family Services (DHFS) summarizing the volume and cost of care provided to all of its residents, Medicaid-funded and otherwise. These cost reports are audited by DHFS staff and form the basis for facility-specific Medicaid daily payment rates for subsequent periods. Medicaid daily payment rate schedules vary annually by facility and by resident classification of level of care within each facility. The data are publicly available. Interested parties can contact the DHFS to request the data; see http://dhfs.wisconsin.gov/contact.htm for contact information.

Utilization of nursing home care is measured in patient days. Medicaid facilities bill the Medicaid fiscal intermediary at the end of each month for total Medicaid patient days incurred in the month, itemized by resident and level of care. Projections of Medicaid patient days by facility and level of care play a key role in the annual process of updating facility Medicaid rate schedules. The projected total patient days are applied to estimated rate schedules to provide an estimate of aggregate Medicaid payments to nursing homes. The rate formula, which translates historical reported costs per patient day for a facility into a proposed rate schedule, is iteratively adjusted until estimated aggregate payments satisfy budget constraints while providing adequate and equitable reimbursement to providers across ownership type, geographic, and other dimensions. Typically DHFS obtains short-term forecasts of one or two fiscal years of aggregate patient days, by separately trending historical Medicaid patient days by level of care. These level-of-care forecasts are stratified by facility roughly in proportion to reported Medicaid total patient days by facility and level of care for the most recently available audited cost report period.

DHFS is interested in determining whether more sophisticated projection techniques exist that might provide more reliable total patient days forecasts by level of care and facility. Because the factors influencing trends in FDD and nursing facility service utilization differ, and nursing facility payments dominate FDD care payments, we focus our analysis initially on nursing facility total patient days.
3. Model Development and Estimation

In this section we use longitudinal data techniques to analyze the nursing home data from State of Wisconsin nursing home cost reports between 1989 and 2001. There are three general classes of longitudinal models that have been discussed in the literature: subject-specific models, marginal models, and transition models (Diggle et al. 2002). In this article, we focus on subject-specific models, which are appropriate for forecasting outcomes at the subject level.

We consider commonly used longitudinal models that assume that the outcome variable follows a multivariate normal distribution. These models are easily implemented and interpreted. We begin in Section 3.1 with a description of the key variables for the entire data set. We then partition the data into an in-sample portion to develop models, and an out-of-sample portion to assess the forecasts of competing models. To calibrate these models and select the most appropriate representation, 4076 observations from 398 nursing facilities, from 1989 to 1999, are included in the in-sample data. Section 3.2 discusses the in-sample model fitting. Section 4 examines the prediction accuracy of different models for the out-of-sample data.

3.1 Summary Statistics

Table 1 describes the variables considered in this analysis. The outcome variable is total patient years (TPY) defined to be number of total patient days in the cost reporting period divided by number of facility operating days in the cost reporting period. The median of total patient years per facility was 86.7 per year. Appendix A describes the statistical decision-making process for choosing to analyze patient years in lieu of patient days. The number of beds and square footage of the nursing home both measure the size of the facility. Not surprisingly, these continuous covariates turn out to be important predictors of TPY; larger facilities have a higher capacity and are likely to have more patients.

Table 1 also describes several categorical covariates. About half of the facilities have self-funding of insurance; these facilities have a higher median TPY than nursing homes that do not self-fund. Approximately 70% of the facilities are Medicare-certified; Medicare-certified facilities are also larger in terms of the median TPY. Regarding the organizational structure, about half (52.4%) are run on a for-profit basis, about one-third (36.8%) are organized as tax exempt, and the remainder are governmental organizations. The government facilities have the highest median TPY. Slightly more than half of the

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Percentage</th>
<th>Median TPY</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPY</td>
<td>Total patient years (median 86.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous covariates:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEAR</td>
<td>Cost report year minus 1988 (1 to 13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NumBed</td>
<td>No. of beds (median 95)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SqrFoot</td>
<td>Nursing home net square footage (in thousands of square feet, median 37.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categorical covariates:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POPID</td>
<td>Nursing home identification number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SelfFundIns</td>
<td>Self-funding of insurance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>50.4</td>
<td>97.1</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td>49.6</td>
<td>74.3</td>
</tr>
<tr>
<td>MCert</td>
<td>Medicare-certified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>70.2</td>
<td>92.0</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td>29.8</td>
<td>70.0</td>
</tr>
<tr>
<td>Organizational Structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro (nursing home is for-profit)</td>
<td></td>
<td>52.4</td>
<td>82.7</td>
</tr>
<tr>
<td>TaxExempt (nursing home is tax exempt)</td>
<td></td>
<td>36.8</td>
<td>86.2</td>
</tr>
<tr>
<td>Govt (nursing home is a governmental unit)</td>
<td></td>
<td>10.8</td>
<td>111.8</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td>53.8</td>
<td>95.5</td>
</tr>
<tr>
<td>Rural</td>
<td></td>
<td>46.2</td>
<td>78.7</td>
</tr>
<tr>
<td>MSA</td>
<td>Metropolitan Statistical Area code (1 to 13, 0 rural)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2
Summary Statistics of Total Patient Years (TPY), by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>No. of Facilities</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>376</td>
<td>102.0</td>
<td>88.3</td>
<td>67.0</td>
<td>15.1</td>
<td>645.0</td>
<td>65.7</td>
</tr>
<tr>
<td>1990</td>
<td>376</td>
<td>103.4</td>
<td>88.5</td>
<td>66.9</td>
<td>14.2</td>
<td>650.6</td>
<td>64.7</td>
</tr>
<tr>
<td>1991</td>
<td>368</td>
<td>104.1</td>
<td>90.1</td>
<td>66.8</td>
<td>13.2</td>
<td>655.1</td>
<td>64.2</td>
</tr>
<tr>
<td>1992</td>
<td>372</td>
<td>105.2</td>
<td>89.3</td>
<td>67.5</td>
<td>16.0</td>
<td>653.0</td>
<td>64.2</td>
</tr>
<tr>
<td>1993</td>
<td>367</td>
<td>105.5</td>
<td>89.7</td>
<td>68.0</td>
<td>16.5</td>
<td>657.3</td>
<td>64.5</td>
</tr>
<tr>
<td>1994</td>
<td>370</td>
<td>104.4</td>
<td>89.4</td>
<td>67.2</td>
<td>14.7</td>
<td>658.6</td>
<td>64.4</td>
</tr>
<tr>
<td>1995</td>
<td>370</td>
<td>104.6</td>
<td>90.4</td>
<td>67.5</td>
<td>16.6</td>
<td>669.3</td>
<td>64.5</td>
</tr>
<tr>
<td>1996</td>
<td>372</td>
<td>101.8</td>
<td>87.8</td>
<td>65.3</td>
<td>15.7</td>
<td>657.3</td>
<td>64.1</td>
</tr>
<tr>
<td>1997</td>
<td>370</td>
<td>99.7</td>
<td>85.3</td>
<td>62.3</td>
<td>15.7</td>
<td>614.4</td>
<td>62.5</td>
</tr>
<tr>
<td>1998</td>
<td>365</td>
<td>97.2</td>
<td>85.1</td>
<td>59.5</td>
<td>11.9</td>
<td>551.5</td>
<td>61.2</td>
</tr>
<tr>
<td>1999</td>
<td>370</td>
<td>92.4</td>
<td>82.7</td>
<td>52.3</td>
<td>11.6</td>
<td>439.6</td>
<td>56.6</td>
</tr>
<tr>
<td>2000</td>
<td>362</td>
<td>88.8</td>
<td>80.5</td>
<td>46.1</td>
<td>11.6</td>
<td>314.7</td>
<td>51.9</td>
</tr>
<tr>
<td>2001</td>
<td>355</td>
<td>89.7</td>
<td>81.1</td>
<td>49.0</td>
<td>12.3</td>
<td>440.7</td>
<td>54.7</td>
</tr>
<tr>
<td>1989–2001</td>
<td>4793</td>
<td>100.0</td>
<td>86.7</td>
<td>62.7</td>
<td>11.6</td>
<td>669.3</td>
<td>62.7</td>
</tr>
</tbody>
</table>

Note: Years 1989–99 correspond to in-sample data; years 2000-2001 correspond to out-of-sample data.

facilities are located in an urban environment (53.8%); these facilities have a higher median TPY than those located in rural environments.

Table 2 shows the distribution of TPY over time. The number of facilities varies from year to year, with an average of 368.7 per year (4,793/13). The number of facilities is relatively stable over time, whereas there has been a small decrease in the typical (mean or median) TPY over time.

Because the standard deviation is large relative to the mean, Table 2 also suggests that the distribution of TPY is skewed. This is reinforced by Figure 1, which provides a histogram of TPY. We see that the distribution is right skewed, with large values of TPY relatively common. The histogram of TPY values on a natural logarithmic scale exhibits a more symmetric, and thinner tailed, distribution.
Table 3  
Correlations among Continuous Covariates

<table>
<thead>
<tr>
<th></th>
<th>LN(TPY)</th>
<th>TPY</th>
<th>NumBed</th>
<th>LN(NumBed)</th>
<th>SqrFoot</th>
<th>LN(SqrFoot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LN(TPY)</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPY</td>
<td>0.905</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NumBed</td>
<td>0.893</td>
<td>0.983</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN(NumBed)</td>
<td>0.983</td>
<td>0.895</td>
<td>0.911</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SqrFoot</td>
<td>0.749</td>
<td>0.825</td>
<td>0.819</td>
<td>0.749</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>LN(SqrFoot)</td>
<td>0.848</td>
<td>0.783</td>
<td>0.781</td>
<td>0.849</td>
<td>0.900</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Motivated by this, we consider the natural logarithm of TPY, LN(TPY), as a potential outcome variable and then convert back to the original units for predicting TPY. Table 3 exhibits correlations among TPY, LN(TPY), and the continuous covariates. This table presents the natural logarithm of number of beds, LN(NumBed), and the natural logarithm of square footage, LN(SqrFoot), and suggests that covariates on a logarithmic scale are more closely related to LN(TPY) than those on the original scale. Therefore, we present our models in terms of continuous covariates on a logarithmic scale.

Knowledge of the size, organizational structure, and other variables will aid in making predictions of TPY. However, for predicting future values of a facility’s TPY, the most important piece of information is past behavior. To underscore the importance of the dynamics, Figure 2 presents multiple time series plots for subsets of facilities. Here LN(TPY) is graphed versus year, with each line connecting annual observations of LN(TPY) for a facility. Only subsets of the facilities have been displayed in these graphs so that they do not appear too cluttered.

Figure 2 underscores the fact that prior values of LN(TPY) are important predictors of future values. Moreover, Figure 2 exhibits the so-called heterogeneity among facilities; observations within a facility

Figure 2  
Multiple Time Series Plots of Logarithmic TPY

(a) Average Total Patient Years < 30  
(b) Average Total Patient Years > 200

Notes: The left-hand panel (a) summarizes a subset of facilities with small average TPY; the right-hand panel (b) corresponds to facilities with large average TPY.
tend to have the same value compared to observations across facilities. Values of TPY are quite stable over time for many facilities; however, some facilities have a substantial amount of variability. We interpret these differences in variability as an aspect of heteroscedasticity. Finally, Figure 2 shows the unbalanced nature of our data; some facilities stopped reporting information within our sampling period, whereas other facilities entered.

### 3.2 In-Sample Model Fitting

The model specification criteria consist of two parts, in-sample selection criteria and out-of-sample assessment criteria. This section focuses on in-sample measures. Section 4 discusses out-of-sample measures.

We include the ordinary regression model as a baseline model to compare with the longitudinal models:

$$\ln(TPY) = \beta_0 + \beta_1 \ln(NumBed) + S_u + \varepsilon_{it}, \quad (3.1)$$

where

$$S_u = \beta_2 \ln(SqrFoot) + \beta_3 Pro_u + \beta_4 \text{TaxExempt}_u + \beta_5 \text{SelfFundIns}_u + \beta_6 \text{MCert}_u + \beta_7 \text{YEAR}_t + \beta_8 \text{YEAR}^2,$$  \quad (3.2)

is a systematic component that is common to each model. This model is also known as a pooled cross-sectional (PCS) model in longitudinal data analysis in that it does not use historical facility-specific information to model the outcome variable.

As suggested by Figure 2, there are strong heterogeneities among the nursing facilities, and strong correlations within each facility over time. This suggests that the PCS model is not appropriate, and longitudinal models are in order. We follow the notation of Frees (2004); for nursing facility $i$ in year $t$, the subject-specific models are of the form

$$y_{it} = z_{it}' \alpha_i + x_{it}' \beta + \varepsilon_{it}, \quad (3.3)$$

where $z_{it} = (z_{it,1}, \ldots, z_{it,q})'$ and $x_{it} = (x_{it,1}, \ldots, x_{it,K})'$ are nonstochastic covariates. The term $\alpha_i$ is a vector of facility-specific parameters corresponding to $z_{it}$, and $\beta$ is a vector of parameters common to all facilities corresponding to $x_{it}$. The vector of error terms is denoted by $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i11})'$. It has mean $(0, \ldots, 0)'$ and $11 \times 11$ temporal variance-covariance matrix $R_i$, where the element in the $r$th row and $s$th column of $R_i$ is $\text{Cov}(y_{ir}, y_{is})$. For facilities with fewer than 11 years of observations, the variance-covariance matrix can be obtained by removing the rows and columns of $R_i$ corresponding to years that are not available.

To account for the heterogeneity among facilities, we use models with facility-specific parameters in the form of equation (3.3). To account for the serial correlation of the outcome variable TPY over time, we assume an autoregressive of order 1 (AR(1)) correlation structure. For an AR(1) correlation structure, we have $R_{rs} = \sigma^2 p^{r-s}$. More complex structures than AR(1) can also be readily fitted; however, as we will see with our data set of only 11 years, AR(1) provides sufficient flexibility. The AR(1) correlation structure makes implicit use of the transition model approach to longitudinal modeling mentioned previously, in that past and current outcomes are assumed to be correlated.

Our first longitudinal representation is a fixed effects (FE) model with facility-specific intercepts that are fixed, unknown parameters. Compared to the PCS model in equation (3.1), this model separates out facility-specific effects that capture much of the time-constant information in the outcome variables; therefore, the estimates of the variability are more precise. Let $\alpha_{i0}$ denote the fixed facility-specific intercept to be estimated. The model is

$$\ln(TPY) = \alpha_{i0} + \beta_1 \ln(NumBed) + S_u + \varepsilon_{it}, \quad (3.4)$$

which we refer to as model FE1.

In addition to allowing intercepts to vary by facility, we can also vary the slope coefficient associated with $\ln(\text{NumBed})$ by facility: that is, we anticipate that the effect of changes of number of beds on
TPY to be facility-specific. We refer to this as model FE2. Let $\alpha_{i1}$ denote the slope coefficient of LN(NumBed) for facility $i$ and use

$$\text{LN(TPY)}_i = \alpha_{i0} + \alpha_{i1}\text{LN(NumBed)}_i + S_{it} + \epsilon_{it}. \quad (3.5)$$

We also consider a situation in which $\alpha_{i}$ is a vector of random variables instead of fixed, unknown parameters, known as “random effects.” Representations that include both fixed and random effects are known as mixed effects (ME) models; our first such model (model ME1) is given in equation (3.6). The form of the model is the same as the FE model with variable intercept in equation (3.4), only the intercept of the $i$th nursing facility is a function of the population intercept $(\beta_0)$ plus a unique contribution from that facility, $\alpha_{i0}$. We assume that $\alpha_{i0}$ are independent of $\epsilon_i$ and are identically and independently distributed with mean 0 and variance $\sigma^2_{\alpha_{0i}}$.

$$\text{LN(TPY)}_i = (\beta_0 + \alpha_{i0}) + \beta_1\text{LN(NumBed)}_i + S_{it} + \text{MSA}_i\gamma + \epsilon_{it}. \quad (3.6)$$

In ME1 we also include the categorical variable MSA representing the metropolitan statistical area. As noted in Table 1, there are 14 different categories represented in this factor, so that 13 covariates $\text{MSA}_i = (\text{MSA}_{i1}, \ldots, \text{MSA}_{i13})'$ and their corresponding fixed regression coefficients $(\gamma = (\beta_2, \ldots, \beta_{21})'$ are included in ME1. The intuition is that facilities in the same MSA share a similar economic climate that can be represented using a constant (within MSA) regression term.

In ME2 we allow the slope of LN(NumBed) to vary by facility $i$; that is, the slope for the $i$th facility is the sum of the population slope and a unique contribution from the facility, $\alpha_{i1}$. As with intercepts, we assume $\alpha_{i1}$ are independent of the error terms and are identically and independently distributed with mean 0 and variance $\sigma^2_{\alpha_{1i}}$. The covariance of $\alpha_{i0}$ and $\alpha_{i1}$ is denoted by $\sigma_{01}$. Model ME2 is given by

$$\text{LN(TPY)}_i = (\beta_0 + \alpha_{i0}) + (\alpha_{i1} + \beta_1)\text{LN(NumBed)}_i + S_{it} + \text{MSA}_i\gamma + \epsilon_{it}. \quad (3.7)$$

Note that in this tutorial article, we include both FE and ME models for illustrative purposes. In practice, analysts often need to decide between these two types of longitudinal models. When there are time-constant covariates within subjects as in our data, such as MSA, ME models are preferred; coefficients for time-constant covariates cannot be estimated in FE models because they are collinear with the subject-specific intercept terms. The sampling method used to select subjects also can aid in the choice between FE and ME models. For example, if the subjects are randomly selected from a population, it is more reasonable to represent the subject-specific effect $\alpha_i$ as a random variable instead of as a fixed, unknown parameter. If, on the other hand, the subjects themselves constitute the population (for example, states in the United States), a FE model is more appropriate. When the choice is not clear, the Hausman test can be employed to choose between FE or ME models (Hausman 1978).

Several popular free or commercial software packages can be used to fit longitudinal models, such as R, Stata, and SAS. We use SAS PROC MIXED to fit both the FE and ME models; illustrative SAS code appears in Appendix B. The variance components are estimated using the restricted maximum likelihood (REML) method rather than the maximum likelihood method. Maximum likelihood estimators of the variance components are often biased and sometimes can be negative. REML produces unbiased, nonnegative estimators (Frees 2004, Chapter 3). Once we estimate the variance components, we can obtain the regression coefficient estimates for the FE or ME models. The estimated variance-covariance matrix of the regression parameters is computed by using the so-called sandwich estimator, which is asymptotically consistent and robust to unsuspected serial correlation and heteroscedasticity (Frees 2004, Chapter 3).

As in ordinary regression models, the statistical significance of individual regression coefficients is determined using t-statistics. Analysts can use penalized likelihood criterion such as Akaike’s Information Criterion (AIC) or Schwarz’s Bayesian Information Criterion (BIC) to compare alternative models (Brockett 1991; Burnham and Anderson 2004). Both statistics include a penalty that is an increasing function of the number of estimated parameters, but do not always lead to the same choice of models. AIC penalizes the number of free parameters less strongly than BIC. (AIC = $-2 \ln$(maximum likelihood) + 2(number of estimable parameters); BIC = $-2 \ln$(maximum likelihood) + ln(number of
subjects) \times (\text{number of estimable parameters}). \) A smaller value of AIC and BIC is associated with a better model fit.

Table 4 summarizes the coefficient estimates. The fit of the PCS model is much poorer than the other models, based on the large value of the AIC and BIC statistics. This result confirms our observation based on the multiple time series plots that the PCS model is not appropriate.

In the other models, the coefficients of the variables LN(SqrFoot), MCert, and YEAR are positively significant; Pro, TaxExempt, and YEAR\(^2\) are negatively significant. Holding the other covariates constant, we interpret the positive coefficient for MCert to mean that Medicare-certified facilities tend to contribute more TPYs. The significant nonzero coefficients on YEAR and YEAR\(^2\) indicate a quadratic trend. TPYs are increasing in the early years of the sample and declining in later years. To illustrate, for the FE1 model, the partial impact of going from year 11 \((0.013 \times 11 - 0.001 \times 11^2 = 0.022)\) to year 12 \((0.013 \times 12 - 0.001 \times 12^2 = 0.012)\) represents an estimated change of \(-0.010\) of logarithmic TPY. One can interpret this as a 1.0% decline going from year 11 (1999) to year 12 (2000).

Comparing models FE1 and FE2, AIC indicates that FE2 provides a better fit, while BIC indicates that FE1 provides a better fit. We note that, for many of the facilities, the number of beds did not change over time. For these facilities, the number of beds is constant and hence perfectly collinear with the constant associated with the intercept term. Thus, slope coefficients are not estimable, and forecasting is not possible for these facilities. Therefore, we drop FE2 from consideration and use the simpler FE1.

Comparing models ME1 and ME2, both AIC and BIC indicate that ME2 provides a better fit. However, one can interpret the correlation terms between the random effects as a correlation; the estimated correlation between \(\alpha_0\) and \(\alpha_1\) is \(-0.165/\sqrt{(0.730) \times (0.037)} = -0.998\). This indicates that the two random effects are highly negatively correlated, and that the slope coefficient is not informative. Therefore, we drop ME2 from consideration and use the simpler ME1.

In FE1 and ME1, the coefficients for both models are largely in agreement in that they provide similar information. Both indicate that our size measures, LN(NumBed) and LN(SqrFoot), are positively statistically significant; the coefficients of Medicare-certified and YEAR are also positively significant; the coefficients of TaxExempt and YEAR\(^2\) are negatively significant; and the coefficient of SelfFundIns is not significant.

<table>
<thead>
<tr>
<th>Regression variables:</th>
<th>Model PCS</th>
<th>Model FE1</th>
<th>Model FE2</th>
<th>Model ME1*</th>
<th>Model ME2*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>(t)-stat</td>
<td>Estimate</td>
<td>(t)-stat</td>
<td>Estimate</td>
</tr>
<tr>
<td>LN(NumBed)</td>
<td>0.950</td>
<td>150.33</td>
<td>0.535</td>
<td>7.82</td>
<td>0.838</td>
</tr>
<tr>
<td>LN(SqrFoot)</td>
<td>0.040</td>
<td>7.05</td>
<td>0.059</td>
<td>2.82</td>
<td>0.027</td>
</tr>
<tr>
<td>Pro</td>
<td>0.021</td>
<td>3.46</td>
<td>-0.104</td>
<td>-2.02</td>
<td>-0.079</td>
</tr>
<tr>
<td>TaxExempt</td>
<td>0.039</td>
<td>6.83</td>
<td>-0.091</td>
<td>-1.85</td>
<td>-0.081</td>
</tr>
<tr>
<td>SelfFundIns</td>
<td>0.002</td>
<td>0.66</td>
<td>0.002</td>
<td>0.51</td>
<td>0.003</td>
</tr>
<tr>
<td>MCert</td>
<td>-0.018</td>
<td>-4.42</td>
<td>0.014</td>
<td>2.95</td>
<td>0.014</td>
</tr>
<tr>
<td>YEAR</td>
<td>0.015</td>
<td>6.93</td>
<td>0.013</td>
<td>4.78</td>
<td>0.009</td>
</tr>
<tr>
<td>YEAR(^2)</td>
<td>-0.001</td>
<td>-6.76</td>
<td>-0.001</td>
<td>-6.79</td>
<td>-0.001</td>
</tr>
<tr>
<td>MSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance Components:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept variance (\sigma_0^2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>Slope variance (\sigma_1^2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept-slope covariance (\sigma_{0,1})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disturbance variance (\sigma^2)</td>
<td>0.010</td>
<td>0.012</td>
<td>0.003</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>Correlation (\rho)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.838</td>
</tr>
<tr>
<td>Goodness-of-fit statistics:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2 residual log-likelihood</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AIC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BIC</td>
</tr>
</tbody>
</table>

\* ME models were estimated including the categorical factor metropolitan statistical area (MSA), although the coefficients are not reported.

Table 4: Logarithmic TPY Model Coefficient Estimates Based on In-Sample Data
4. **Model Validation**

We compare the predictive abilities of our candidate models using the out-of-sample data from the 12th and 13th years, 2000 and 2001. There are 717 observations from 366 nursing facilities in these years. An additional four facilities started after 1999; they are not included in the out-of-sample data.

For each model we used the in-sample data to estimate coefficients and used the estimated coefficients and covariates for the out-of-sample data to compute predicted LN(TPY) (Frees 2004 Chapter 4). Then we exponentiated to get $\hat{TPY}_{i,t+L}$ for each facility $i$ and forecast leads $L = 1, 2$ (corresponding to $t = 12, 13,$ or years 2000, 2001).

We summarize the model accuracy of the forecasts through two statistics, the Mean Absolute Error (MAE),

$$\text{MAE} = \frac{1}{n_{12} + n_{13}} \sum_{L=1}^{2} \sum_{i=1}^{n_{11+L}} |\hat{TPY}_{i,11+L} - TPY_{i,11+L}|,$$

and the Mean Absolute Percentage Error (MAPE),

$$\text{MAPE} = \frac{100}{n_{12} + n_{13}} \sum_{L=1}^{2} \sum_{i=1}^{n_{11+L}} \left|\frac{\hat{TPY}_{i,11+L} - TPY_{i,11+L}}{TPY_{i,11+L}}\right|.$$

Here $n_t$ denotes the number of facilities in year $t$.

Table 5 summarizes the comparison of the models based on the out-of-sample criteria. Table 5 also summarizes forecasts based on the in-sample average of each facility’s TPY: this naive estimator performed poorly. However, an alternative naive estimator, the most recent observation, performed even better than the PCS model on the out-of-sample fit, without any adjustments for trend. The most recent observation is included in our comparisons as a baseline measure; analysts should always consider this naive measure. Although not reported in detail here, this estimator performed poorly in the in-sample assessment. Moreover, the performance of the most recent observation deteriorates rapidly as the forecast horizon ($L$) increases.

As expected, the PCS model performed poorly compared to the other candidate models. Comparing FE1 and ME1, we see that both models performed better than the most recent observation, with the slight edge to ME1.

To underscore the appropriateness of the natural logarithm transformations, we also fit the PCS, FE1, and ME1 models using the original units of the continuous variables TPY, NumBed, and SqrFoot. The forecast results are reported in Table 5. Compared to the models with LN(TPY) as the outcome variable, the FE1 and ME1 models for TPY have larger MAE and MAPE values, indicating that the LN(TPY) models fit the data better.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Absolute Error (MAE)</th>
<th>Mean Absolute Percentage Error (MAPE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most recent observation</td>
<td>4.62</td>
<td>5.47</td>
</tr>
<tr>
<td>Outcome variable LNTPY:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-sample average</td>
<td>12.12</td>
<td>13.36</td>
</tr>
<tr>
<td>PCS</td>
<td>6.05</td>
<td>7.16</td>
</tr>
<tr>
<td>FE1</td>
<td>4.51</td>
<td>5.28</td>
</tr>
<tr>
<td>ME1</td>
<td>4.18</td>
<td>4.84</td>
</tr>
<tr>
<td>Outcome variable TPY:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCS</td>
<td>5.96</td>
<td>7.71</td>
</tr>
<tr>
<td>FE1</td>
<td>5.34</td>
<td>7.02</td>
</tr>
<tr>
<td>ME1</td>
<td>4.45</td>
<td>5.51</td>
</tr>
</tbody>
</table>
When developing forecasting models, it is prudent to test the ability of the model to predict over a range of time periods to establish robustness. To assess model performance at different points in time, Table 6 reports the forecast results on the logarithm transformed TPY of another in-sample period, beginning in 1989 and ending in 1997. Similar to the results in Table 5, both the FE1 and ME1 models outperform the other candidates. Interestingly, the PCS model forecasts better than the most recent observation. Because the most recent observation is based on only observation per facility, it is highly variable. Thus, it is not surprising that its forecasting performance fluctuates wildly over different time periods.

5. SUMMARY AND CONCLUDING REMARKS

This article illustrated predictive modeling using State of Wisconsin nursing home data. We emphasized that predictive modeling is a process that involves problem identification, data analysis, and candidate model development, estimation, and validation. Our predictive modeling approach involved longitudinal modeling and compared three types of models, the PCS, FE, and ME models, with two simple approaches, the in-sample average and most recent observation. Introducing covariates and allowing for either a fixed or random intercept term by facility improved the prediction, compared to the simpler approaches or an ordinary regression model. Although not surprising, this result is significant given the common industry practice of using cross-sectional algorithms for predictive modeling. These models are easily computed with appropriate software packages such as Stata, R, or in our case, SAS.

As mentioned in the Introduction, predictive modeling has been used for provider profiling, provider reimbursement, and identification of high-cost users. In all of these applications, costs can and should be linked over time, whether they be by physician, by organization, or by individual. Longitudinal modeling of costs over time accounts for the heterogeneity of individuals, through inclusion of individual-specific intercept and slope coefficients.

The longitudinal data approach used in this article is only one of several predictive modeling approaches. Other approaches include continuance tables, multiple regression analysis, generalized linear models (GLMs), two-part models, Bayesian analysis, finite mixtures of distributions, and statistical algorithms such as clustering, principal components, classification and regression trees, multivariate adaptive regression splines, and neural networks.

Continuance tables traditionally have been used in the health insurance industry for the predictive modeling of frequency distributions for outcomes such as length of stay or claim duration. For example, our Wisconsin nursing home data could have been analyzed by developing a continuance table for total patient days or years, using historical data and actuarial assumptions to estimate the probability that a resident would spend \( N \) days in a nursing home, where \( N \) is random count variable. Such an approach could be useful for analyzing discrete outcome variables. However, continuance tables cannot be easily generalized to continuous outcome variables, such as the expected cost of treating a nursing home resident or claim severities. There are many articles in the literature that discuss continuance tables; some references are provided in Waters and Phil (1989).

Multiple regression analysis is one of the most widely used predictive modeling techniques in health care. Some widely used multiple regression predictive models for determining provider reimbursement

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Absolute Error (MAE)</th>
<th>Mean Absolute Percentage Error (MAPE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-sample average</td>
<td>11.64</td>
<td>12.27</td>
</tr>
<tr>
<td>Most recent observation</td>
<td>6.72</td>
<td>7.25</td>
</tr>
<tr>
<td>PCS</td>
<td>5.62</td>
<td>6.35</td>
</tr>
<tr>
<td>FE1</td>
<td>4.85</td>
<td>5.31</td>
</tr>
<tr>
<td>ME1</td>
<td>4.86</td>
<td>5.16</td>
</tr>
</tbody>
</table>
are the Diagnostic Cost Group Hierarchical Condition Category models (DCG/HCC) (Ash et al. 2000). These models incorporate patient diagnosis codes as covariates to aid in the prediction of provider payments. Other papers contain discussions of health-care-related multiple regression models, including Ash and Byrne-Logan (1998), Fowles et al. (1996), Pope et al. (2000), Pope et al. (2004), Sales et al. (2003), and Zhao et al. (2005).

The GLM approach is a popular method for modeling skewed data. GLMs are flexible in that the link and variance functions used in this method can account for data issues such as overdispersion (Frees 2004, Chapter 10). Papers that have discussed GLMs with regard to predicting health care costs and utilization include Blough, Madden, and Hornbrook (1999), Diehr et al. (1999), Daniels and Gatsonis (1999), and Manning and Mullahy (2001).

Two-part models estimate the utilization of health care separately from the cost. The first component models the likelihood that a patient will use any medical services using either a logistic or probit regression model. The second component models the amount of medical services utilized by the patient, conditional on the patient having used any services. If multiple regression analysis is used, then the outcome variable is typically the logarithm of health care cost, whereas if a GLM is used to model this component, then no transformation is needed. The product of the expected value of each component provides the expected health care costs for an individual (Blough et al. 1999; Deb and Trivedi 2002; Diehr et al. 1999).

Other possible approaches to model health care costs include Bayesian predictive modeling (De Alba 2002; Fellingham, Tolley, and Herzog 2005; Verrall 2004) or the use of finite mixture distributions (Deb and Burgess 2003; Deb and Holmes 2000; Deb and Trivedi 1997). Clustering, principal components, classification and regression trees, multivariate adaptive regression splines, and neural networks are more complex models that can be readily implemented using specialized algorithms embedded in statistical computer software. These techniques allow for high flexibility in the specification of the regression function and easily allow for complex transformations of and interactions between covariates (Berry and Linoff 2004; Hastie, Tibshirani, and Friedman 2001).

Health care modeling using nonnormal distributions can be effective in reducing prediction error. In the future we will extend this work to examine the data using nonnormal distributions. Also, the predictive power of models determined by the predictive modeling process depends on the quality of the data used to generate the models. Analysts often face problems with missing data, which can arise in longitudinal data as attrition, where subjects that contributed at least one subject-time observation fail to provide observations in other time periods. Depending on the type of nonresponse, parameters of predictive models may be improperly estimated because of potential selection bias. A related concern pertains to potential endogeneity of model covariates. For example, the analyst may be unable to include all covariates that are related to the outcome variable in the theoretical model because of data limitations; this is commonly known in the statistical literature as omitted variable bias, which can also cause model parameters to be biased. Future work will develop methods to allow analysts to account for these potential issues in predictive modeling.

**APPENDIX A**

**Diagnostic Analyses**

As with all statistical modeling exercises, the models introduced in Section 3 were the result of an intense study of the data and the literature supporting health care utilization. Although this article focuses on the steps ending with the in-sample estimation and the out-of-sample validation, in any data analysis there are many complex decisions that need to be made prior to recommendation of these candidate models.

In regression and longitudinal data analysis, the process of examining a preliminary model fit and using information about any lack of fit to improve the model specification is known as “diagnostic analysis.” Residual analysis is an important type of diagnostic analysis. Residuals are defined as the
observed value minus the fitted value, and the analysis involves both the numeric and graphical inspection of the estimated models (Frees 2004).

To illustrate this process, it is interesting to consider the outcome variable total patient years (TPY). When first examining the data, policymakers most interested in the cost reports worked with the total of patient days that a nursing home facility utilized in a cost report period without adjustments to the length of the period. For this reason, we originally used LN(Total Patient Days) as the outcome variable without adjustment of the number of facility operating days in the cost reporting period (DaysOpen).

We had already determined that this variable, like TPY, was quite skewed and required a natural logarithmic transform. To get a sense of the data, about 2% of observations were not operating for the whole cost report year. Four observations’ cost report periods were longer than one calendar year; this can occur if a facility changes its fiscal year or some other significant change in a facility’s characteristics occurs, such as a change in ownership.

Residual analysis uncovered some important patterns that we had not adequately addressed in our models LN(Total Patient Days) as the outcome variable. Figure 3 shows the residual analysis of the mixed effects (ME1) model, using the same covariates as reported in Table 4. (The results of the pooled cross-sectional (PCS) and fixed effects (FE1) models are similar, thus not shown here.) The left panel in Figure 3 exhibits the residuals versus the fitted value of the ME1 model, and the right panel shows the relationship between the residuals and length of cost report period (DaysOpen). It is clear that the residuals decrease as the number of facility operating days decreases, indicated by a positive relation between these variables. One possibility is to include DaysOpen as a covariate in the regression. However, we chose to define a new variable, TPY, as number of total patient days in the cost reporting period divided by number of facility operating days in the cost report period to take the length of cost report period into consideration. This selection also accounts for some heteroscedasticity that is not evident in Figure 3.
APPENDIX B

ILLUSTRATIVE SAS CODE FOR MODEL ESTIMATION

```sas
proc format;
value yesno 0='1:no' 1='0:yes';
run;

proc mixed data=INSAMPLE noclprint order=formatted;
title 'Pooled Cross-Sectional Model (Model PCS)';
class POPID MSA;
model LNTPY=MSA LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
   /s outp=cs noint;
   format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
title 'Fixed Variable Intercept Model (Model FE1)';
class POPID;
model LNTPY=POPID LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
   /s noint outp=fe1;
   repeated /type=ar(1) sub=POPID r;
   format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
title 'Fixed Variable Intercept and Slope Model (Model FE2)';
class POPID;
model LNTPY=POPID POPID*LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
   /s noint outp=fe2;
   repeated /type=ar(1) sub=POPID r rcorr;
   format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
title 'Error Components Model (Model ME1)';
class POPID MSA;
model LNTPY=MSA LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
   /s outp=me1 noint;
   random intercept /sub=POPID g;
   repeated /type=ar(1) sub=POPID r;
   format PRO TaxExempt SelfFundIns MCert yesno.;
run;

proc mixed data=INSAMPLE noclprint empirical order=formatted;
title 'Random Variable Intercept and Slope Model (Model ME2)';
class POPID MSA;
model LNTPY=MSA LNNumBed LNSqrFoot PRO TaxExempt SelfFundIns MCert YEAR|YEAR
   /s outp=me2 noint;
   random intercept LNNumBed/ sub=POPID type=UN g;
   repeated / type=ar(1) sub=POPID r;
   format PRO TaxExempt SelfFundIns MCert yesno.;
run;
```

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation, Grant Number SES-0436274, and the Agency for Healthcare Research and Quality, Grant Number R03 HS16519, and the Assurant Health Professorship in Actuarial Science.
REFERENCES


Discussions on this paper can be submitted until January 1, 2008. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.
<table>
<thead>
<tr>
<th>Category</th>
<th>Variable Description</th>
<th>Percent of data</th>
<th>Average of Positive Expend</th>
<th>Percent of Positive Expend</th>
<th>Average of Pos Expend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>Age in years between 18 to 65 (mean: 39.0)</td>
<td>52.7</td>
<td>0.91</td>
<td>10.7</td>
<td>8.53</td>
</tr>
<tr>
<td>GENDER</td>
<td>1 if female</td>
<td>47.3</td>
<td>0.40</td>
<td>4.7</td>
<td>8.66</td>
</tr>
<tr>
<td>GENDER</td>
<td>1 if male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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>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></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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>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></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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 physical health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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>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>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 mental health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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 limitation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>0 if otherwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>77.7</td>
</tr>
<tr>
<td>Income compared to poverty line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>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>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>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 coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>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>100.0</td>
<td>0.67</td>
<td>7.9</td>
<td>8.32</td>
</tr>
</tbody>
</table>
Table 16.2. *Comparison of OLS, Tobit MLE and Two-Stage Estimates*

<table>
<thead>
<tr>
<th>Effect</th>
<th>Parameter Estimate</th>
<th>Parameter $t$-ratio</th>
<th>Parameter Estimate</th>
<th>Parameter $t$-ratio</th>
<th>Parameter Estimate</th>
<th>Parameter $t$-ratio*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.123</td>
<td>-0.525</td>
<td>-33.016</td>
<td>-8.233</td>
<td>2.760</td>
<td>0.617</td>
</tr>
<tr>
<td>AGE</td>
<td>0.001</td>
<td>0.091</td>
<td>-0.006</td>
<td>-0.118</td>
<td>0.001</td>
<td>0.129</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.379</td>
<td>3.711</td>
<td>5.727</td>
<td>4.107</td>
<td>0.271</td>
<td>1.617</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.115</td>
<td>-0.459</td>
<td>-1.732</td>
<td>-0.480</td>
<td>-0.091</td>
<td>-0.480</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.054</td>
<td>0.365</td>
<td>0.025</td>
<td>0.015</td>
<td>0.043</td>
<td>0.262</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.350</td>
<td>0.726</td>
<td>3.745</td>
<td>0.723</td>
<td>0.250</td>
<td>0.445</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>1.702</td>
<td>3.828</td>
<td>1.849</td>
<td>0.203</td>
<td>1.065</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.255</td>
<td>1.693</td>
<td>3.459</td>
<td>1.790</td>
<td>0.196</td>
<td>1.143</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.146</td>
<td>1.133</td>
<td>1.805</td>
<td>1.056</td>
<td>0.117</td>
<td>0.937</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>-0.014</td>
<td>-0.089</td>
<td>0.628</td>
<td>0.329</td>
<td>-0.024</td>
<td>-0.149</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>-0.027</td>
<td>-0.209</td>
<td>-0.030</td>
<td>-0.019</td>
<td>-0.026</td>
<td>-0.202</td>
</tr>
<tr>
<td>POOR</td>
<td>2.297</td>
<td>7.313</td>
<td>13.352</td>
<td>4.436</td>
<td>1.780</td>
<td>1.810</td>
</tr>
<tr>
<td>FAIR</td>
<td>-0.001</td>
<td>-0.004</td>
<td>1.354</td>
<td>0.528</td>
<td>-0.014</td>
<td>-0.068</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.188</td>
<td>1.346</td>
<td>2.740</td>
<td>1.480</td>
<td>0.143</td>
<td>1.018</td>
</tr>
<tr>
<td>VGOOD</td>
<td>0.084</td>
<td>0.622</td>
<td>1.506</td>
<td>0.815</td>
<td>0.063</td>
<td>0.533</td>
</tr>
<tr>
<td>MNHPOOR</td>
<td>0.000</td>
<td>-0.001</td>
<td>-0.482</td>
<td>-0.211</td>
<td>-0.011</td>
<td>-0.041</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.415</td>
<td>3.103</td>
<td>4.695</td>
<td>3.000</td>
<td>0.306</td>
<td>1.448</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.482</td>
<td>-2.716</td>
<td>-6.575</td>
<td>-3.035</td>
<td>-0.338</td>
<td>-1.290</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.309</td>
<td>-1.868</td>
<td>-4.359</td>
<td>-2.241</td>
<td>-0.210</td>
<td>-0.952</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.175</td>
<td>-0.976</td>
<td>-3.414</td>
<td>-1.619</td>
<td>-0.099</td>
<td>-0.438</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.116</td>
<td>-0.478</td>
<td>-2.274</td>
<td>-0.790</td>
<td>-0.065</td>
<td>-0.243</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.594</td>
<td>4.486</td>
<td>8.534</td>
<td>4.130</td>
<td>0.455</td>
<td>2.094</td>
</tr>
</tbody>
</table>

Inverse Mill’s Ratio $\lambda$ | -3.616 | -0.642 |
Scale $\sigma^2$ | 4.999 | 14.738 | 4.997 |

* Two-stage $t$-ratios are calculated using heteroscedasticity-consistent standard errors.
Table 16.3. *Comparison of Full and Reduced Two-Part Models*

<table>
<thead>
<tr>
<th>Effect</th>
<th>Full Model</th>
<th></th>
<th></th>
<th>Reduced Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Severity</td>
<td></td>
<td>Frequency</td>
<td>Severity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>$t$-ratio</td>
<td>Estimate</td>
<td>$t$-ratio</td>
<td>Estimate</td>
<td>$t$-ratio</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.001</td>
<td>-0.154</td>
<td>0.012</td>
<td>1.368</td>
<td>0.020</td>
<td>2.437</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.395</td>
<td>4.176</td>
<td>-0.104</td>
<td>-0.469</td>
<td>0.395</td>
<td>4.178</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.108</td>
<td>-0.429</td>
<td>-0.397</td>
<td>-0.641</td>
<td>-0.108</td>
<td>-0.427</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.008</td>
<td>0.062</td>
<td>0.088</td>
<td>0.362</td>
<td>0.009</td>
<td>0.073</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.284</td>
<td>0.778</td>
<td>-0.639</td>
<td>-0.905</td>
<td>0.285</td>
<td>0.780</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>1.958</td>
<td>-0.649</td>
<td>-2.035</td>
<td>0.281</td>
<td>1.950</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.239</td>
<td>1.765</td>
<td>0.016</td>
<td>0.052</td>
<td>0.237</td>
<td>1.754</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.132</td>
<td>1.099</td>
<td>-0.078</td>
<td>-0.294</td>
<td>0.130</td>
<td>1.085</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.048</td>
<td>0.356</td>
<td>-0.597</td>
<td>-2.066</td>
<td>0.049</td>
<td>0.362</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>0.002</td>
<td>0.017</td>
<td>-0.415</td>
<td>-1.745</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td>POOR</td>
<td>0.955</td>
<td>4.576</td>
<td>0.597</td>
<td>1.594</td>
<td>0.939</td>
<td>4.805</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.087</td>
<td>0.486</td>
<td>-0.211</td>
<td>-0.527</td>
<td>0.079</td>
<td>0.450</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.184</td>
<td>1.422</td>
<td>0.145</td>
<td>0.502</td>
<td>0.182</td>
<td>1.412</td>
</tr>
<tr>
<td>VGOOD</td>
<td>0.095</td>
<td>0.736</td>
<td>0.373</td>
<td>1.233</td>
<td>0.094</td>
<td>0.728</td>
</tr>
<tr>
<td>MNHPOOR</td>
<td>-0.027</td>
<td>-0.164</td>
<td>-0.176</td>
<td>-0.579</td>
<td>-0.177</td>
<td>-0.640</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.318</td>
<td>2.941</td>
<td>0.235</td>
<td>0.981</td>
<td>0.311</td>
<td>3.022</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.468</td>
<td>-3.131</td>
<td>0.490</td>
<td>1.531</td>
<td>-0.470</td>
<td>-3.224</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.314</td>
<td>-2.318</td>
<td>0.472</td>
<td>1.654</td>
<td>-0.314</td>
<td>-2.345</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.241</td>
<td>-1.626</td>
<td>0.550</td>
<td>1.812</td>
<td>-0.241</td>
<td>-1.633</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.145</td>
<td>-0.716</td>
<td>0.067</td>
<td>0.161</td>
<td>-0.146</td>
<td>-0.721</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.580</td>
<td>4.154</td>
<td>1.293</td>
<td>3.944</td>
<td>0.579</td>
<td>4.147</td>
</tr>
</tbody>
</table>

Scale $\sigma^2$: 1.249

Scale $\sigma^2$: 1.333
<table>
<thead>
<tr>
<th>Effect</th>
<th>Negative Binomial Frequency Parameter Estimate</th>
<th>Negative Binomial Severity Parameter estimate</th>
<th>Ordinary Regression Severity Parameter Estimate</th>
<th>Gamma Regression Severity Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interception</td>
<td>-4.214</td>
<td>-9.169</td>
<td>7.424</td>
<td>15.514</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.005</td>
<td>-0.756</td>
<td>-0.006</td>
<td>-0.747</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.617</td>
<td>3.351</td>
<td>-0.385</td>
<td>-1.952</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.153</td>
<td>-0.306</td>
<td>-0.340</td>
<td>-0.588</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.144</td>
<td>0.639</td>
<td>0.146</td>
<td>0.686</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.445</td>
<td>0.634</td>
<td>-0.331</td>
<td>-0.465</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.492</td>
<td>1.683</td>
<td>-0.547</td>
<td>-1.792</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.619</td>
<td>2.314</td>
<td>0.303</td>
<td>1.070</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.391</td>
<td>1.603</td>
<td>0.108</td>
<td>0.424</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.023</td>
<td>0.089</td>
<td>-0.789</td>
<td>-2.964</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>-0.085</td>
<td>-0.399</td>
<td>-0.722</td>
<td>-3.396</td>
</tr>
<tr>
<td>POOR</td>
<td>1.927</td>
<td>5.211</td>
<td>0.664</td>
<td>1.964</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.226</td>
<td>0.627</td>
<td>-0.188</td>
<td>-0.486</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.385</td>
<td>1.483</td>
<td>0.223</td>
<td>0.802</td>
</tr>
<tr>
<td>VGOOD</td>
<td>0.348</td>
<td>1.349</td>
<td>0.429</td>
<td>1.511</td>
</tr>
<tr>
<td>MNHPOOR</td>
<td>-0.177</td>
<td>-0.583</td>
<td>-0.221</td>
<td>-0.816</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.714</td>
<td>3.499</td>
<td>0.579</td>
<td>2.720</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.622</td>
<td>-2.139</td>
<td>0.723</td>
<td>2.517</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.482</td>
<td>-1.831</td>
<td>0.720</td>
<td>2.768</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.460</td>
<td>-1.611</td>
<td>0.631</td>
<td>2.241</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.465</td>
<td>-1.131</td>
<td>-0.056</td>
<td>-0.135</td>
</tr>
<tr>
<td>INSURE</td>
<td>1.312</td>
<td>4.207</td>
<td>1.500</td>
<td>4.551</td>
</tr>
<tr>
<td>Dispersion</td>
<td>2.177</td>
<td>1.314</td>
<td>1.131</td>
<td>1.131</td>
</tr>
</tbody>
</table>
Regression Modeling with Actuarial and Financial Applications

Edward W. Frees
This book will be published by Cambridge University Press in January, 2010
Additional information is at the website
http://research3.bus.wisc.edu/RegActuaries

copyright
©Edward W. Frees 2010
Reprinted with permission
Contents

11 Categorical Dependent Variables

11.1 Binary Dependent Variables 3
11.2 Logistic and Probit Regression Models 5
  11.2.1 Using Nonlinear Functions of Explanatory Variables 5
  11.2.2 Threshold Interpretation 7
  11.2.3 Random Utility Interpretation 7
  11.2.4 Logistic Regression 8
11.3 Inference for Logistic and Probit Regression Models 9
  11.3.1 Parameter Estimation 9
  11.3.2 Additional Inference 10
11.4 Application: Medical Expenditures 12
11.5 Nominal Dependent Variables 15
  11.5.1 Generalized Logit 16
  11.5.2 Multinomial Logit 18
  11.5.3 Nested Logit 21
11.6 Ordinal Dependent Variables 22
  11.6.1 Cumulative Logit 22
  11.6.2 Cumulative Probit 24
11.7 Further Reading and References 24
11.8 Exercises 25
11.9 Technical Supplements - Likelihood-Based Inference 31
  11.9.1 Properties of Likelihood Functions 32
  11.9.2 Maximum Likelihood Estimators 33
  11.9.3 Hypothesis Tests 34
  11.9.4 Information Criteria 35

16 Frequency-Severity Models

16.1 Introduction 37
16.2 Tobit Model 38
16.3 Application: Medical Expenditures 40
16.4 Two-Part Model 43
16.5 Aggregate Loss Model 46
16.6 Further Reading and References 48
16.7 Exercises 51
11
Categorical Dependent Variables

Chapter Preview. A model with a categorical dependent variable allows one to predict whether an observation is a member of a distinct group, or category. Binary variables represent an important special case; they can indicate whether or not an event of interest has occurred. In actuarial and financial applications, the event may be whether a claim occurs, a person purchases insurance, a person retires or a firm becomes insolvent. The chapter introduces logistic regression and probit models of binary dependent variables. Categorical variables may also represent more than two groups, known as multcategory outcomes. Multicategory variables may be unordered or ordered, depending on whether it makes sense to rank the variable outcomes. For unordered outcomes, known as nominal variables, the chapter introduces generalized logits and multinomial logit models. For ordered outcomes, known as ordinal variables, the chapter introduces cumulative logit and probit models.

11.1 Binary Dependent Variables
We have already introduced binary variables as a special type of discrete variable that can be used to indicate whether or not a subject has a characteristic of interest, such as gender for a person or ownership of a captive insurance company for a firm. Binary variables also describe whether or not an event of interest has occurred, such as an accident. A model with a binary dependent variable allows one to predict whether an event has occurred or a subject has a characteristic of interest.

Example: MEPS Expenditures. Section 11.4 will describe an extensive database from the Medical Expenditure Panel Survey (MEPS) on hospitalization utilization and expenditures. For these data, we will consider

\[ y_i = \begin{cases} 
1 & \text{ith person was hospitalized during the sample period} \\
0 & \text{otherwise} 
\end{cases} \]

There are \( n = 2,000 \) persons in this sample, distributed as:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not hospitalized ( y = 0 )</td>
<td>902 (95.3%)</td>
<td>941 (89.3%)</td>
</tr>
<tr>
<td>Hospitalized ( y = 1 )</td>
<td>44 (4.7%)</td>
<td>113 (10.7%)</td>
</tr>
<tr>
<td>Total</td>
<td>946</td>
<td>1,054</td>
</tr>
</tbody>
</table>

Table 11.1. Hospitalization by Gender
Table 11.1 suggests that gender has an important influence on whether someone becomes hospitalized.

Like the linear regression techniques introduced in prior chapters, we are interested in using characteristics of a person, such as their age, sex, education, income and prior health status, to help explain the dependent variable \( y \). Unlike the prior chapters, now the dependent variable is discrete and not even approximately normally distributed. In limited circumstances, linear regression can be used with binary dependent variables - this application is known as a linear probability model.

**Linear probability models**

To introduce some of the complexities encountered with binary dependent variables, denote the probability that the response equals 1 by \( \pi_i = \Pr(y_i = 1) \). A binary random variable has a Bernoulli distribution. Thus, we may interpret the mean response as the probability that the response equals one, that is, \( \mathbb{E}(y_i) = 0 \times \Pr(y_i = 0) + 1 \times \Pr(y_i = 1) = \pi_i \). Further, the variance is related to the mean through the expression \( \text{Var}(y_i) = \pi_i(1 - \pi_i) \).

We begin by considering a linear model of the form

\[
y_i = x_i' \beta + \varepsilon_i,
\]

known as a linear probability model. Assuming \( \mathbb{E}(\varepsilon_i) = 0 \), we have that \( \mathbb{E}(y_i) = x_i' \beta = \pi_i \). Because \( y_i \) has a Bernoulli distribution, \( \text{Var}(y_i) = x_i' \beta (1 - x_i' \beta) \). Linear probability models are used because of the ease of parameter interpretations. For large data sets, the computational simplicity of ordinary least squares estimators is attractive when compared to some complex alternative nonlinear models introduced later in this chapter. As described in Chapter 3, ordinary least squares estimators for \( \beta \) have desirable properties. It is straightforward to check that the estimators are consistent and asymptotically normal under mild conditions on the explanatory variables \( \{x_i\} \). However, linear probability models have several drawbacks that are serious for many applications.

**Drawbacks of the Linear Probability Model**

- **Fitted values can be poor.** The expected response is a probability and thus must vary between 0 and 1. However, the linear combination, \( x_i' \beta \), can vary between negative and positive infinity. This mismatch implies, for example, that fitted values may be unreasonable.

- **Heteroscedasticity.** Linear models assume homoscedasticity (constant variance) yet the variance of the response depends on the mean that varies over observations. The problem of varying variability is known as heteroscedasticity.

- **Residual analysis is meaningless.** The response must be either a 0 or 1 although the regression models typically regards distribution of the error term as continuous. This mismatch implies, for example, that the usual residual analysis in regression modeling is meaningless.
To handle the heteroscedasticity problem, a (two-stage) weighted least squares procedure is possible. In the first stage, one uses ordinary least squares to compute estimates of $\beta$. With this estimate, an estimated variance for each subject can be computed using the relation $\text{Var} \ y_i = x_i'\hat{\beta}(1 - x_i'\hat{\beta})$. At the second stage, a weighted least squares is performed using the inverse of the estimated variances as weights to arrive at new estimates of $\beta$. It is possible to iterate this procedure, although studies have shown that there are few advantages in doing so (see Carroll and Ruppert, 1988). Alternatively, one can use ordinary least squares estimators of $\beta$ with standard errors that are robust to heteroscedasticity (see Section 5.7.2).

11.2 Logistic and Probit Regression Models

11.2.1 Using Nonlinear Functions of Explanatory Variables

To circumvent the drawbacks of linear probability models, we consider alternative models in which we express the expectation of the response as a function of explanatory variables, $\pi_i = \pi(x_i'\beta) = \Pr(y_i = 1|x_i)$. We focus on two special cases of the function $\pi(\cdot)$:

- $\pi(z) = \frac{1}{1 + \exp(-z)} = \frac{e^z}{1 + e^z}$, the logit case, and
- $\pi(z) = \Phi(z)$, the probit case.

Here, $\Phi(\cdot)$ is the standard normal distribution function. The choice of the identity function (a special kind of linear function), $\pi(z) = z$, yields the linear probability model. In contrast, $\pi$ is nonlinear for both the logit and probit cases. These two functions are similar in that they are almost linearly related over the interval $0.1 \leq p \leq 0.9$. Thus, to a large extent, the function choice is dependent on the preferences of the analyst. Figure 11.1 compares the logit and probit functions showing that it will be difficult to distinguish between the two specifications with most data sets.

The inverse of the function, $\pi^{-1}$, specifies the form of the probability that is linear in the explanatory variables, that is, $\pi^{-1}(\pi_i) = x_i'\beta$. In Chapter 13, we refer to this inverse as the link function.

Fig. 11.1. Comparison of Logit and Probit (Standard Normal) Distribution Functions
Example: Credit Scoring. Banks, credit bureaus and other financial institutions develop “credit scores” for individuals that are used to predict the likelihood that the borrower will repay current and future debts. Individuals who do not meet stipulated repayment schedules in a loan agreement are said to be in “default.” A credit score is then a predicted probability of being in default, with the credit application providing the explanatory variables used in developing the credit score. The choice of explanatory variables depends on the purpose of the application; credit scoring is used for issuing credit cards for making small consumer purchases as well as mortgage applications for multimillion dollar houses. In Table 11.2, Hand and Henley (1997) provide a list of typical characteristics that are used in credit scoring.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Potential Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time at present address</td>
<td>0-1, 1-2, 3-4, 5+ years</td>
</tr>
<tr>
<td>Home status</td>
<td>Owner, tenant, other</td>
</tr>
<tr>
<td>Postal Code</td>
<td>Band A, B, C, D, E</td>
</tr>
<tr>
<td>Telephone</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Applicant’s annual income</td>
<td>£(0-10000), £(10,000-20,000) £(20,000+)</td>
</tr>
<tr>
<td>Credit card</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Type of bank account</td>
<td>Check and/or savings, none</td>
</tr>
<tr>
<td>Age</td>
<td>18-25, 26-40, 41-55, 55+ years</td>
</tr>
<tr>
<td>County Court judgements</td>
<td>Number</td>
</tr>
<tr>
<td>Type of occupation</td>
<td>Coded</td>
</tr>
<tr>
<td>Purpose of loan</td>
<td>Coded</td>
</tr>
<tr>
<td>Marital status</td>
<td>Married, divorced, single, widow, other</td>
</tr>
<tr>
<td>Time with bank</td>
<td>Years</td>
</tr>
<tr>
<td>Time with employer</td>
<td>Years</td>
</tr>
</tbody>
</table>

Source: Hand and Henley (1997)

With credit application information and default experience, a logistic regression model can be used to fit the probability of default with credit scores resulting from fitted values. Wiginton (1980) provides an early application of logistic regression to consumer credit scoring. At that time, other statistical methods known as discriminant analysis were at the cutting edge of quantitative scoring methodologies. In their review article, Hand and Henley (1997) discuss other competitors to logistic regression including machine learning systems and neural networks. As noted by Hand and Henley, there is no uniformly “best” method. Regression techniques are important in their own right due to their widespread usage and because they can provide a platform for learning about newer methods.

Credit scores provide estimates of the likelihood of defaulting on loans but issuers of credit are also interested in the amount and timing of debt repayment. For example, a “good” risk may repay a credit balance so promptly that little profit is earned by the lender. Further, a “poor” mortgage risk may default on a loan so late in the duration of the contract that a sufficient profit was earned by the lender. See Gourieroux and Jasiak (2007) for a broad discussion of how credit modeling can be used to assess the riskiness and profitability of loans.
11.2 Logistic and Probit Regression Models

11.2.2 Threshold Interpretation

Both the logit and probit cases can be interpreted as follows. Suppose that there exists an underlying linear model, \( y_i^* = x_i' \beta + \varepsilon_i^* \). Here, we do not observe the response \( y_i^* \) yet interpret it to be the “propensity” to possess a characteristic. For example, we might think about the financial strength of an insurance company as a measure of its propensity to become insolvent (no longer capable of meeting its financial obligations). Under the threshold interpretation, we do not observe the propensity but we do observe when the propensity crosses a threshold. It is customary to assume that this threshold is 0, for simplicity. Thus, we observe

\[
 y_i = \begin{cases} 
 0 & y_i^* \leq 0 \\
 1 & y_i^* > 0 
\end{cases}.
\]

To see how the logit case is derived from the threshold model, assume a logistic distribution function for the disturbances, so that

\[
 Pr(\varepsilon_i^* \leq a) = \frac{1}{1 + \exp(-a)}.
\]

Like the normal distribution, one can verify by calculating the density that the logistic distribution is symmetric about zero. Thus, \( -\varepsilon_i^* \) has the same distribution as \( \varepsilon_i^* \) and so

\[
 \pi_i = Pr(y_i = 1|x_i) = Pr(y_i^* > 0) = Pr(\varepsilon_i^* \leq x_i' \beta) = \frac{1}{1 + \exp(-x_i' \beta)} = \pi(x_i' \beta).
\]

This establishes the threshold interpretation for the logit case. The development for the probit case is similar and is omitted.

11.2.3 Random Utility Interpretation

Both the logit and probit cases are also justified by appealing to the following “random utility” interpretation of the model. In some economic applications, individuals select one of two choices. Here, preferences among choices are indexed by an unobserved utility function; individuals select the choice that provides the greater utility.

For the \( i \)th subject, we use the notation \( u_i \) for this utility function. We model the utility \( U \) as a function of an underlying value \( V \) plus random noise \( \varepsilon \), that is, \( U_{ij} = u_i(V_{ij} + \varepsilon_{ij}) \), where \( j \) may be 1 or 2, corresponding to the choice. To illustrate, we assume that the individual chooses the category corresponding to \( j = 1 \) if \( U_{i1} > U_{i2} \) and denote this choice as \( y_i = 1 \). Assuming that \( u_i \) is a strictly increasing function, we have

\[
 Pr(y_i = 1) = Pr(U_{i2} < U_{i1}) = Pr(u_i(V_{i2} + \varepsilon_{i2}) < u_i(V_{i1} + \varepsilon_{i1})) = Pr(\varepsilon_{i2} - \varepsilon_{i1} < V_{i1} - V_{i2}).
\]

To parameterize the problem, assume that the value \( V \) is an unknown linear combination of explanatory variables. Specifically, we take \( V_{i2} = 0 \) and \( V_{i1} = x_i' \beta \). We may take the difference in the errors, \( \varepsilon_{i2} - \varepsilon_{i1} \), as normal or logistic, corresponding to the probit and logit cases, respectively. The logistic distribution is satisfied if the errors are assumed to have an extreme-value, or Gumbel, distribution (see, for example, Amemiya, 1985).
11.2.4 Logistic Regression

An advantage of the logit case is that it permits closed-form expressions, unlike the normal distribution function. Logistic regression is another phrase used to describe the logit case.

Using $p = \pi(z) = (1 + e^{-z})^{-1}$, the inverse of $\pi$ is calculated as $z = \pi^{-1}(p) = \ln(p/(1 - p))$. To simplify future presentations, we define

$$\text{logit}(p) = \ln \left( \frac{p}{1 - p} \right)$$

to be the logit function. With a logistic regression model, we represent the linear combination of explanatory variables as the logit of the success probability, that is, $x'_i \beta = \text{logit}(\pi_i)$.

**Odds interpretation**

When the response $y$ is binary, knowing only $p = \Pr(y = 1)$ summarizes the entire distribution. In some applications, a simple transformation of $p$ has an important interpretation. The lead example of this is the odds, given by $p/(1 - p)$. For example, suppose that $y$ indicates whether or not a horse wins a race and $p$ is the probability of the horse winning. If $p = 0.25$, then the odds of the horse winning is $0.25/(1 - 0.25) = 0.3333$. We might say that the odds of winning are 0.3333 to 1, or one to three. Equivalently, we say that the probability of not winning is $1 - p = 0.75$ so that the odds of the horse not winning is $0.75/(1 - 0.75) = 3$ and the odds against the horse are three to one.

Odds have a useful interpretation from a betting standpoint. Suppose that we are playing a fair game and that we place a bet of $1$ with one to three odds. If the horse wins, then we get our $1$ back plus winnings of $3$. If the horse loses, then we lose our bet of $1$. It is a fair game in the sense that the expected value of the game is zero because we win $3$ with probability $p = 0.25$ and lose $1$ with probability $1 - p = 0.75$. From an economic standpoint, the odds provide the important numbers (bet of $1$ and winnings of $3$), not the probabilities. Of course, if we know $p$, then we can always calculate the odds. Similarly, if we know the odds, we can always calculate the probability $p$.

The logit is the logarithmic odds function, also known as the log odds.

**Odds ratio interpretation**

To interpret the regression coefficients in the logistic regression model, $\beta = (\beta_0, \ldots, \beta_k)'$, we begin by assuming that $j$th explanatory variable, $x_{ij}$, is either 0 or 1. Then, with the notation $x_i = (x_{i0}, \ldots, x_{ij}, \ldots, x_{ik})'$, we may interpret

$$\beta_j = \left( x_{i0}, \ldots, 1, \ldots, x_{ik} \right)' \beta - \left( x_{i0}, \ldots, 0, \ldots, x_{ik} \right)' \beta$$

$$= \ln \left( \frac{\Pr(y_i = 1| x_{ij} = 1)}{1 - \Pr(y_i = 1| x_{ij} = 1)} \right) - \ln \left( \frac{\Pr(y_i = 1| x_{ij} = 0)}{1 - \Pr(y_i = 1| x_{ij} = 0)} \right)$$

Thus,

$$e^{\beta_j} = \frac{\Pr(y_i = 1| x_{ij} = 1)}{\Pr(y_i = 1| x_{ij} = 0)} / \frac{1 - \Pr(y_i = 1| x_{ij} = 1)}{1 - \Pr(y_i = 1| x_{ij} = 0)}.$$
This shows that $e^{\bar{\beta}_j}$ can be expressed as the ratio of two odds, known as the odds ratio. That is, the numerator of this expression is the odds when $x_{ij} = 1$, whereas the denominator is the odds when $x_{ij} = 0$. Thus, we can say that the odds when $x_{ij} = 1$ are $\exp(\bar{\beta}_j)$ times as large as the odds when $x_{ij} = 0$. To illustrate, suppose $\beta_j = 0.693$, so that $\exp(\bar{\beta}_j) = 2$. From this, we say that the odds (for $y = 1$) are twice as great for $x_{ij} = 1$ as for $x_{ij} = 0$.

Similarly, assuming that $j$th explanatory variable is continuous (differentiable), we have

$$
\beta_j = \frac{\partial}{\partial x_{ij}} \ln \left( \frac{\Pr(y_i = 1|x_{ij})}{1 - \Pr(y_i = 1|x_{ij})} \right) = \frac{\partial}{\partial x_{ij}} \Pr(y_i = 1|x_{ij}) / (1 - \Pr(y_i = 1|x_{ij})) \Pr(y_i = 1|x_{ij}) / (1 - \Pr(y_i = 1|x_{ij})).
$$

(11.1)

Thus, we may interpret $\beta_j$ as the proportional change in the odds ratio, known as an elasticity in economics.

---

**Example: MEPS Expenditures - Continued.** Table 11.1 shows that the percentage of females who were hospitalized is 10.7%; alternatively, the odds of females being hospitalized is $0.107/(1-0.107) = 0.120$. For males, the percentage is 4.7% so that the odds were 0.0493. The odds ratio is 0.120/0.0493 = 2.434; females are more than twice as likely to be hospitalized as males.

From a logistic regression fit (described in Section 11.4), the coefficient associated with gender is 0.733. Based on this model, we say that females are $\exp(0.733) = 2.081$ times as likely as males to be hospitalized. The regression estimate of the odds ratio controls for additional variables (such as age and education) compared to the basic calculation based on raw frequencies.

---

### 11.3 Inference for Logistic and Probit Regression Models

#### 11.3.1 Parameter Estimation

The customary method of estimation for logistic and probit models is maximum likelihood, described in further detail in Section 11.9. To provide intuition, we outline the ideas in the context of binary dependent variable regression models.

The likelihood is the observed value of the probability function. For a single observation, the likelihood is

$$
\begin{cases}
1 - \pi_i & \text{if } y_i = 0 \\
\pi_i & \text{if } y_i = 1
\end{cases}
$$

The objective of maximum likelihood estimation is to find the parameter values that produce the largest likelihood. Finding the maximum of the logarithmic function yields the same solution as finding the maximum of the corresponding function. Because it is generally computationally simpler, we consider the logarithmic (or log-) likelihood, written as

$$
\begin{cases}
\ln (1 - \pi_i) & \text{if } y_i = 0 \\
\ln \pi_i & \text{if } y_i = 1
\end{cases}
$$

(11.2)
More compactly, the log-likelihood of a single observation is
\[ y_i \ln \pi(x'_i \beta) + (1 - y_i) \ln (1 - \pi(x'_i \beta)), \]

where \( \pi_i = \pi(x'_i \beta). \) Assuming independence among observations, the likelihood of the data set is a product of likelihoods of each observation. Taking logarithms, the log-likelihood of the data set is the sum of log-likelihoods of single observations.

The log-likelihood of the data set is
\[ L(\beta) = \sum_{i=1}^{n} \{ y_i \ln \pi(x'_i \beta) + (1 - y_i) \ln (1 - \pi(x'_i \beta)) \}, \quad (11.3) \]

The log-likelihood is viewed as a function of the parameters, with the data held fixed. In contrast, the joint probability mass function is viewed as a function of the realized data, with the parameters held fixed.

The method of maximum likelihood involves finding the values of \( \beta \) that maximize the log-likelihood. The customary method of finding the maximum is taking partial derivatives with respect to the parameters of interest and finding roots of the resulting equations. In this case, taking partial derivatives with respect to \( \beta \) yields the score equations
\[ \frac{\partial}{\partial \beta} L(\beta) = \sum_{i=1}^{n} x_i (y_i - \pi(x'_i \beta)) \frac{\pi'(x'_i \beta)}{\pi(x'_i \beta)(1 - \pi(x'_i \beta))} = 0, \quad (11.4) \]

where \( \pi' \) is the derivative of \( \pi. \) The solution of these equations, denoted as \( \mathbf{b}_{MLE}, \) is the maximum likelihood estimator. For the logit function the score equations reduce to
\[ \frac{\partial}{\partial \beta} L(\beta) = \sum_{i=1}^{n} x_i (y_i - \pi(x'_i \beta)) = 0, \quad (11.5) \]

where \( \pi(z) = 1/(1 + \exp(-z)). \)

11.3.2 Additional Inference

An estimator of the large sample variance of \( \beta \) may be calculated taking partial derivatives of the score equations. Specifically, the term
\[ I(\beta) = -E \left( \frac{\partial^2}{\partial \beta \partial \beta} L(\beta) \right) \]

is the information matrix. As a special case, using the logit function and equation (11.5), straightforward calculations show that the information matrix is
\[ I(\beta) = \sum_{i=1}^{n} \sigma_i^2 x_i x'_i \]

where \( \sigma_i^2 = \pi(x'_i \beta)(1 - \pi(x'_i \beta)). \) The square root of the \((j + 1)st\) diagonal element of this matrix evaluated at \( \beta = \mathbf{b}_{MLE} \) yields the standard error for \( b_{j,MLE}, \) denoted as \( se(b_{j,MLE}). \)
11.3 Inference for Logistic and Probit Regression Models

To assess the overall model fit, it is customary to cite likelihood ratio test statistics in nonlinear regression models. To test the overall model adequacy $H_0 : \beta = 0$, we use the statistic

$$LRT = 2 \times (L(b_{MLE}) - L_0),$$

where $L_0$ is the maximized log-likelihood with only an intercept term. Under the null hypothesis $H_0$, this statistic has a chi-square distribution with $k$ degrees of freedom. Section 11.9.3 describes likelihood ratio test statistics in greater technical detail.

As described in Section 11.9, measures of goodness of fit can be difficult to interpret in nonlinear models. One measure is the so-called max−scaled $R^2$, defined as $R_{ms}^2 = R^2 / R_{max}^2$, where

$$R^2 = 1 - \left( \frac{\exp(L_0/n)}{\exp(L(b_{MLE})/n)} \right)$$

and $R_{max}^2 = 1 - \exp(L_0/n)^2$. Here, $L_0/n$ represents the average value of this log-likelihood.

Another measure is a “pseudo-$R^2$”

$$\frac{L(b_{MLE}) - L_0}{L_{max} - L_0},$$

where $L_0$ and $L_{max}$ is the log-likelihood based on only an intercept and on the maximum achievable, respectively. Like the coefficient of determination, the pseudo-$R^2$ takes on values between zero and one, with larger values indicating a better fit to the data. Other versions of pseudo-$R^2$’s are available in the literature, see, for example, Cameron and Trivedi (1998). An advantage of this pseudo-$R^2$ measure is its link to hypothesis testing of regression coefficients.

**Example: Job Security.** Valletta (1999) studied declining job security using the Panel Survey of Income Dynamics (PSID) database. We consider here one of the regression models presented by Valletta, based on a sample of male heads of households that consists of $n = 24,168$ observations over the years 1976-1992, inclusive. The PSID survey records reasons why men left their most recent employment, including plant closures, “quit” and changed jobs for other reasons. However, Valletta focused on dismissals (“laid off” or “fired”) because involuntary separations are associated with job insecurity.

Table 11.3 presents a probit regression model run by Valletta (1999), using dismissals as the dependent variable. In addition to the explanatory variables listed in Table 11.3, other variables controlled for consisted of education, marital status, number of children, race, years of full-time work experience and its square, union membership, government employment, logarithmic wage, the U.S. employment rate and location as measured through the Metropolitan Statistical Area residence. In Table 11.3, tenure is years employed at the current firm. Further, sector employment was measured by examining the Consumer Price Survey employment in 387 sectors of the economy, based on 43 industry categories and nine regions of the country.

On the one hand, the tenure coefficient reveals that more experienced workers are less likely to be dismissed. On the other hand, the coefficient associated with the
interaction between tenure and time trend reveals an increasing dismissal rate for experienced workers.

The interpretation of the sector employment coefficients is also of interest. With an average tenure of about 7.8 years in the sample, we see the low tenure men are relatively unaffected by changes in sector employment. However, for more experienced men, there is an increasing probability of dismissal associated with sectors of the economy where growth declines.

Table 11.3. Dismissal Probit Regression Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure</td>
<td>-0.084</td>
<td>0.010</td>
</tr>
<tr>
<td>Time Trend</td>
<td>-0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Tenure*(Time Trend)</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>Change in Logarithmic Sector Employment</td>
<td>0.094</td>
<td>0.057</td>
</tr>
<tr>
<td>Tenure*(Change in Logarithmic Sector Employment)</td>
<td>-0.020</td>
<td>0.009</td>
</tr>
</tbody>
</table>

-2 Log Likelihood: 7,027.8
Pseudo-$R^2$: 0.097

11.4 Application: Medical Expenditures

This section considers data from the Medical Expenditure Panel Survey (MEPS), conducted by the U.S. Agency of Health Research and Quality. MEPS is a probability survey that provides nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian population. This survey collects detailed information on individuals and each medical care episode by type of services including physician office visits, hospital emergency room visits, hospital outpatient visits, hospital inpatient stays, all other medical provider visits, and use of prescribed medicines. This detailed information allows one to develop models of health care utilization to predict future expenditures. We consider MEPS data from the first panel of 2003 and take a random sample of $n = 2,000$ individuals between ages 18 and 65.

Dependent Variable

Our dependent variable is an indicator of positive expenditures for inpatient admissions. For MEPS, inpatient admissions include persons who were admitted to a hospital and stayed overnight. In contrast, outpatient events include hospital outpatient department visits, office-based provider visits and emergency room visits excluding dental services. (Dental services, compared to other types of health care services, are more predictable and occur on a more regular basis.) Hospital stays with the same date of admission and discharge, known as “zero-night stays,” were included in outpatient counts and expenditures. Payments associated with emergency room visits that immediately preceded an inpatient stay were included in the inpatient expenditures. Prescribed medicines that can be linked to hospital admissions were included in inpatient expenditures (not in outpatient utilization).
11.4 Application: Medical Expenditures

Explanatory Variables

Explanatory variables that can help explain health care utilization are categorized as demographic, geographic, health status, education and economic factors. Demographic factors include age, sex and ethnicity. As persons age, the rate at which their health deteriorates increases with age; as a result, age has an increasing impact on the demand for health care. Sex and ethnicity can be treated as proxies for inherited health and social habits in maintaining health. For a geographic factor, we use region to proxy the accessibility of health care services and the overall economic or regional impact on residents’ health care behavior.

The demand for medical services is thought to be influenced by individuals’ health status and education. In MEPS, self-rated physical health, mental health and any functional or activity related limitations during the sample period are used as proxies for health status. Education tends to have ambiguous impact on the demand for health care services. One theory is that more educated persons are more aware of health risks, thus being more active in maintaining their health; as a result, educated persons may be less prone to severe diseases leading to hospital admissions. Another theory is that less educated persons have greater exposure to health risks and, through exposure, develop a greater tolerance for certain types of risks. In MEPS, education is proxied by degrees received and categorized into three different levels: lower than high school, high school, and college or above education.

Economic covariates include income and insurance coverage. A measure of income in MEPS is income relative to the poverty line. This approach is appropriate because it summarizes effects of different levels of income on health care utilization in constant dollars. Insurance coverage is also an important variable in explaining health care utilization. One issue with health insurance coverage is that it reduces the out-of-pocket prices paid by insureds and thus induces moral hazard. Research associated with the Rand Health Insurance Experiment empirically suggested that cost sharing effects from insurance coverage will affect primarily the number of medical contacts rather than the intensity of each contact. This motivated our introduction of a binary variable that takes the value of 1 if a person had any public or private health insurance for at least one month, and 0 otherwise.

Summary Statistics

Table 11.4 describes these explanatory variables and provides summary statistics that suggest their effects on the probability of positive inpatient expenditures. For example, we see that females had a higher overall utilization than males. Specifically, 10.7% of females had a positive expenditure during the year compared to only 4.7% for males. Similarly, utilizations vary by other covariates, suggesting their importance as predictors of expenditures.
Table 11.4. Percent of Positive Expenditures by Explanatory Variable

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Description</th>
<th>Percent of data</th>
<th>Percent Positive Exp</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>
</tr>
<tr>
<td></td>
<td>GENDER</td>
<td>1 if female</td>
<td>52.7</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 if male</td>
<td>47.3</td>
<td>4.7</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>ASIAN</td>
<td>1 if Asian</td>
<td>4.3</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>BLACK</td>
<td>1 if Black</td>
<td>14.8</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>NATIVE</td>
<td>1 if Native</td>
<td>1.1</td>
<td>13.6</td>
</tr>
<tr>
<td></td>
<td>WHITE</td>
<td>Reference level</td>
<td>79.9</td>
<td>7.5</td>
</tr>
<tr>
<td>Region</td>
<td>NORTHEAST</td>
<td>1 if Northeast</td>
<td>14.3</td>
<td>10.1</td>
</tr>
<tr>
<td></td>
<td>MIDWEST</td>
<td>1 if Midwest</td>
<td>19.7</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td>SOUTH</td>
<td>1 if South</td>
<td>38.2</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>WEST</td>
<td>Reference level</td>
<td>27.9</td>
<td>5.4</td>
</tr>
<tr>
<td>Education</td>
<td>COLLEGE</td>
<td>1 if college or higher degree</td>
<td>27.2</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>HIGHSCHOOL</td>
<td>1 if high school degree</td>
<td>43.3</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reference level is lower than high school degree</td>
<td>29.5</td>
<td>8.8</td>
</tr>
<tr>
<td>Self-rated</td>
<td>POOR</td>
<td>1 if poor</td>
<td>3.8</td>
<td>36.0</td>
</tr>
<tr>
<td>physical health</td>
<td>FAIR</td>
<td>1 if fair</td>
<td>9.9</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>GOOD</td>
<td>1 if good</td>
<td>29.9</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>VGOOD</td>
<td>1 if very good</td>
<td>31.1</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>MNHPoor</td>
<td>Reference level is excellent health</td>
<td>25.4</td>
<td>5.1</td>
</tr>
<tr>
<td>Self-rated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mental health</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ANYLIMIT</td>
<td>1 if any functional/activity limitation</td>
<td>22.3</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 if otherwise</td>
<td>77.7</td>
<td>5.9</td>
</tr>
<tr>
<td>Income</td>
<td>HINCOME</td>
<td>1 if high income</td>
<td>31.6</td>
<td>5.4</td>
</tr>
<tr>
<td>compared to</td>
<td>MINCOME</td>
<td>1 if middle income</td>
<td>29.9</td>
<td>7.0</td>
</tr>
<tr>
<td>poverty line</td>
<td>LINCOME</td>
<td>1 if low income</td>
<td>15.8</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>NPOOR</td>
<td>1 if near poor</td>
<td>5.8</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reference level is poor/negative</td>
<td>17.0</td>
<td>13.0</td>
</tr>
<tr>
<td>Insurance</td>
<td>INSURE</td>
<td>1 if covered by public/private health insurance in any month of 2003</td>
<td>77.8</td>
<td>9.2</td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>0 if have no health insurance in 2003</td>
<td>22.3</td>
<td>3.1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>100.0</td>
<td>7.9</td>
</tr>
</tbody>
</table>

Table 11.5 summarizes the fit of several binary regression models. Fits are reported under the “Full Model” column for all variables using the logit function. The \( t \)-ratios for many of the explanatory variables exceed two in absolute value, suggesting that they are useful predictors. From an inspection of these \( t \)-ratios, one might consider a more parsimonious model by removing statistically insignificant variables. Table 11.5 shows a “Reduced Model,” where the age and mental health status variables have been removed. To assess their joint significance, we can compute a likelihood ratio test statistic as twice the change in the log-likelihood. This turns out to be only \( 2 \times (−488.78 − (−488.69)) = 0.36 \). Comparing this to a chi-square distribution with \( df = 2 \) degrees of freedom results in a \( p \)-value= 0.835, indicating that the additional parameters for age and mental health status are not statistically significant. Table 11.5 also provides probit model fits. Here, we see that the results are similar to the logit model fits, according to sign of the coefficients and their significance, suggesting that for this application there is little difference in the two specifications.
11.5 Nominal Dependent Variables

We now consider a response that is an unordered categorical variable, also known as a nominal dependent variable. We assume that the dependent variable $y$ may take on values $1, 2, \ldots, c$, corresponding to $c$ categories. When $c > 2$, we refer to the data as “multicategory,” also known as polychotomous or polytomous.

In many applications, the response categories correspond to an attribute possessed or choices made by individuals, households or firms. Some applications include:

- employment choice, such as Valletta (1999)
- mode of transportation, such as the classic work by McFadden (1978)
- type of health insurance, as in Browne and Frees (2007).

For an observation from subject $i$, denote the probability of choosing the $j$th category as $\pi_{ij} = \Pr(y_i = j)$, so that $\pi_1 + \cdots + \pi_c = 1$. In general, we will model these probabilities as a (known) function of parameters and use maximum likelihood estimation for statistical inference. Let $y_{ij}$ be a binary variable that is 1 if $y_i = j$. Extending equation (11.2) to $c$ categories, the likelihood for the $i$th subject is:

\[
\begin{array}{c|c|c|c|c|c}
\text{Effect} & \text{Full Model} & \text{Reduced Model} & \text{Reduced Model} \\
\hline
\text{Intercept} & -4.239 & -8.982 & -4.278 & -10.094 & -2.281 & -11.432 \\
\text{AGE} & -0.001 & -0.180 & 0.004 & 0.019 & 0.009 & 0.073 \\
\text{GENDER} & 0.733 & 3.812 & 0.732 & 3.806 & 0.395 & 4.178 \\
\text{ASIAN} & -0.219 & -0.411 & -0.219 & -0.412 & -0.108 & -0.427 \\
\text{BLACK} & -0.001 & -0.003 & 0.004 & 0.019 & 0.009 & 0.073 \\
\text{NATIVE} & 0.610 & 0.926 & 0.612 & 0.930 & 0.285 & 0.780 \\
\text{NORTHEAST} & 0.609 & 2.112 & 0.604 & 2.098 & 0.281 & 1.950 \\
\text{MIDWEST} & 0.524 & 1.904 & 0.517 & 1.883 & 0.237 & 1.754 \\
\text{SOUTH} & 0.339 & 1.376 & 0.328 & 1.342 & 0.130 & 1.085 \\
\text{COLLEGE} & 0.068 & 0.255 & 0.070 & 0.263 & 0.049 & 0.362 \\
\text{HIGHSCHOOL} & 0.004 & 0.017 & 0.009 & 0.041 & 0.003 & 0.030 \\
\text{POOR} & 1.712 & 4.385 & 1.652 & 4.575 & 0.939 & 4.805 \\
\text{FAIR} & 0.136 & 0.375 & 0.109 & 0.306 & 0.079 & 0.450 \\
\text{GOOD} & 0.376 & 1.429 & 0.368 & 1.405 & 0.182 & 1.412 \\
\text{VGGOOD} & 0.178 & 0.667 & 0.174 & 0.655 & 0.094 & 0.728 \\
\text{MNHPOOR} & -0.113 & -0.369 & & & \\
\text{ANYLIMIT} & 0.564 & 2.680 & 0.545 & 2.704 & 0.311 & 3.022 \\
\text{HINCOME} & -0.921 & -3.101 & -0.919 & -3.162 & -0.470 & -3.224 \\
\text{MINCOME} & -0.609 & -2.315 & -0.604 & -2.317 & -0.314 & -2.345 \\
\text{LINCOME} & -0.411 & -1.453 & -0.408 & -1.449 & -0.241 & -1.633 \\
\text{NPOOR} & -0.201 & -0.528 & -0.204 & -0.534 & -0.146 & -0.721 \\
\text{INSURE} & 1.234 & 4.047 & 1.227 & 4.031 & 0.579 & 4.147 \\
\hline
\text{Log-Likelihood} & -488.69 & -488.78 & -486.98 \\
\text{AIC} & 1,021.38 & 1,017.56 & 1,013.96 \\
\end{array}
\]
Categorical Dependent Variables

\[ \prod_{j=1}^{c} (\pi_{i,j})^{y_{i,j}} = \begin{cases} \pi_{i,1} & \text{if } y_{i} = 1 \\ \pi_{i,2} & \text{if } y_{i} = 2 \\ \vdots & \vdots \\ \pi_{i,c} & \text{if } y_{i} = c \end{cases} \]

Thus, assuming independence among observations, the total log-likelihood is

\[ L = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{i,j} \ln \pi_{i,j}. \]

With this framework, standard maximum likelihood estimation is available (Section 11.9). Thus, our main task is to specify an appropriate form for \( \pi \).

11.5.1 Generalized Logit

Like standard linear regression, generalized logit models employ linear combinations of explanatory variables of the form:

\[ V_{i,j} = \mathbf{x}_{i}^{'} \beta_{j}. \] (11.6)

Because the dependent variables are not numerical, we cannot model the response \( y \) as a linear combination of explanatory variables plus an error. Instead we use the probabilities

\[ \Pr(y_{i} = j) = \pi_{i,j} = \frac{\exp(V_{i,j})}{\sum_{k=1}^{c} \exp(V_{i,k})}. \] (11.7)

Note here that \( \beta_{j} \) is the corresponding vector of parameters that may depend on the alternative \( j \) whereas the explanatory variables \( \mathbf{x}_{i} \) do not. So that probabilities sum to one, a convenient normalization for this model is \( \beta_{c} = 0 \). With this normalization and the special case of \( c = 2 \), the generalized logit reduces to the logit model introduced in Section 11.2.

Parameter interpretations

We now describe an interpretation of coefficients in generalized logit models, similar to the logistic model. From equations (11.6) and (11.7), we have

\[ \ln \frac{\Pr(y_{i} = j)}{\Pr(y_{i} = c)} = V_{i,j} - V_{i,c} = \mathbf{x}_{i}^{'} \beta_{j}. \]

The left-hand side of this equation is interpreted to be the logarithmic odds of choosing choice \( j \) compared to choice \( c \). Thus, we may interpret \( \beta_{j} \) as the proportional change in the odds ratio.

Generalized logits have an interesting nested structure that we will explore briefly in Section 11.5.3. That is, it is easy to check that, conditional on not choosing the first category, the form of \( \Pr(y_{i} = j|y_{i} \neq 1) \) has a generalized logit form in equation (11.7). Further, if \( j \) and \( h \) are different alternatives, we note that
\[ \Pr(y_i = j | y_i = j \text{ or } y_i = h) = \frac{\Pr(y_i = j)}{\Pr(y_i = j) + \Pr(y_i = h)} = \frac{\exp(V_{i,j})}{\exp(V_{i,j}) + \exp(V_{i,h})} = \frac{1}{1 + \exp(x_i' (\bar{\beta}_h - \bar{\beta}_j))}. \]

This has a logit form that was introduced in Section 11.2.

**Special Case - Intercept only model.** To develop intuition, we now consider the model with only intercepts. Thus, let \( x_i = 1 \) and \( \beta_j = \beta_{0,j} = \alpha_j \). With the convention \( \alpha_c = 0 \), we have

\[ \Pr(y_i = j) = \pi_{i,j} = \frac{e^{\alpha_j}}{e^{\alpha_1} + e^{\alpha_2} + \cdots + e^{\alpha_{c-1}} + 1} \]

and

\[ \ln \frac{\Pr(y_i = j)}{\Pr(y_i = c)} = \alpha_j. \]

From the second relation, we may interpret the \( j \)th intercept \( \alpha_j \) to be the logarithmic odds of choosing alternative \( j \) compared to alternative \( c \).

**Example: Job Security - Continued.** This is a continuation of the Section 11.2 example on the determinants of job turnover, based on the work of Valetta (1999). The first analysis of this data considered only the binary dependent variable dismissal as this outcome is the main source of job insecurity. Valetta (1999) also presented results from a generalized logit model, his primary motivation being that the economic theory describing turnover implies that other reasons for leaving a job may affect dismissal probabilities.

For the generalized logit model, the response variable has \( c = 5 \) categories: dismissal, left job because of plant closures, “quit,” changed jobs for other reasons and no change in employment. The “no change in employment” category is the omitted one in Table 11.6. The explanatory variables of the generalized logit are the same as the probit regression; the estimates summarized in Table 11.1 are reproduced here for convenience.

Table 11.6 shows that turnover declines as tenure increases. To illustrate, consider a typical man in the 1992 sample where we have \( t = 16 \) and focus on dismissal probabilities. For this value of time, the coefficient associated with tenure for dismissal is \(-0.221 + 16 (0.008) = -0.093 \) (due to the interaction term). From this, we interpret an additional year of tenure to imply that the dismissal probability is \( \exp(-0.093) = 91\% \) of what it would be otherwise, representing a decline of 9%.

Table 11.6 also shows that the generalized coefficients associated with dismissal are similar to the probit fits.

The standard errors are also qualitatively similar, although higher for the generalized logits when compared to the probit model. In particular, we again see that the coefficient associated with the interaction between tenure and time trend reveals an increasing dismissal rate for experienced workers. The same is true for the rate of quitting.
11.5.2 Multinomial Logit

Similar to equation (11.6), an alternative linear combination of explanatory variables is

\[ V_{i,j} = x'_{i,j} \beta, \]

where \( x_{i,j} \) is a vector of explanatory variables that depends on the \( j \)th alternative whereas the parameters \( \beta \) do not. Using the expressions in equations (11.7) and (11.8) forms the basis of the multinomial logit model, also known as the conditional logit model (McFadden, 1974). With this specification, the total log-likelihood is

\[
L = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{i,j} \ln \pi_{i,j} = \sum_{i=1}^{n} \left[ \sum_{j=1}^{c} y_{i,j} x'_{i,j} \beta - \ln \sum_{k=1}^{c} \exp(x'_{i,k} \beta) \right].
\]

This straightforward expression for the likelihood enables maximum likelihood inference to be easily performed.

The generalized logit model is a special case of the multinomial logit model. To see this, consider explanatory variables \( x_i \) and parameters \( \beta_j \), each of dimension \( k \times 1 \). Define

\[
\begin{pmatrix}
0 \\
\vdots \\
0
\end{pmatrix}
\quad \text{and} \quad
\beta =
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_c
\end{pmatrix}.
\]
Specifically, \( x_{i,j} \) is defined as \( j - 1 \) zero vectors (each of dimension \( k \times 1 \)), followed by \( x_i \) and then followed by \( c - j \) zero vectors. With this specification, we have \( x_{i,j}' \beta = x_{i}' \beta_j \). Thus, a statistical package that performs multinomial logit estimation can also perform generalized logit estimation through the appropriate coding of explanatory variables and parameters. Another consequence of this connection is that some authors use the descriptor multinomial logit when referring to the generalized logit model.

Moreover, through similar coding schemes, multinomial logit models can also handle linear combinations of the form:

\[
V_i = x_{i,1,j}' \beta + x_{i,2,j}' \beta_j.
\]

Here, \( x_{i,1,j} \) are explanatory variables that depend on the alternative whereas \( x_{i,2} \) do not. Similarly, \( \beta_j \) are parameters that depend on the alternative whereas \( \beta \) do not. This type of linear combination is the basis of a mixed logit model. As with conditional logits, it is customary to choose one set of parameters as the baseline and specify \( \beta_c = 0 \) to avoid redundancies.

To interpret parameters for the multinomial logit model, we may compare alternatives \( h \) and \( k \) using equations (11.7) and (11.8), to get

\[
\ln \frac{\Pr(y_i = h)}{\Pr(y_i = k)} = (x_{i,h} - x_{i,k})' \beta.
\]

Thus, we may interpret \( \beta_j \) as the proportional change in the odds ratio, where the change is the value of the \( j \)th explanatory variable, moving from the \( k \)th to the \( h \)th alternative.

With equation (11.7), note that \( \pi_{i,1}/\pi_{i,2} = \exp(V_{i,1})/\exp(V_{i,2}) \). This ratio does not depend on the underlying values of the other alternatives, \( V_{i,j} \), for \( j = 3, \ldots, c \). This feature, called the independence of irrelevant alternatives, can be a drawback of the multinomial logit model for some applications.

**Example: Choice of Health Insurance.** To illustrate, Browne and Frees (2007) examined \( c = 4 \) health insurance choices, consisting of:

- \( y = 1 \) - an individual covered by group insurance,
- \( y = 2 \) - an individual covered by private, non-group insurance,
- \( y = 3 \) - an individual covered by government, but not private insurance or
- \( y = 4 \) - an individual not covered by health insurance.

Their data on health insurance coverage came from the March supplement of the Current Population Survey (CPS), conducted by the Bureau of Labor Statistics. Browne and Frees (2007) analyzed approximately 10,800 single person households per year, covering 1988-1995, yielding \( n = 86,475 \) observations. They examined whether underwriting restrictions, laws passed to prohibit insurers from discrimination, facilitate or discourage consumption of health insurance. They focused on disability laws that prohibited insurers from using physical impairment (disability) as an underwriting criterion.

Table 11.7 suggests that disability laws have little effect on the average health insurance purchasing behavior. To illustrate, for individuals surveyed with disability laws in effect, 57.6% purchased group health compared to 59.3% of those where
restrictions were not in effect. Similarly, 19.9% were uninsured when disability restrictions were in effect compared to 20.1% when they were not. In terms of odds, when disability restrictions were in effect, the odds of purchasing group health insurance compared to becoming uninsured are 57.6/19.9 = 2.895. When disability restrictions were not in effect, the odds are 2.946. The odds ratio, 2.895/2.946 = 0.983, indicates that there is little change in the odds when comparing whether or not disability restrictions were in effect.

Table 11.7. Percentages of Health Coverage by Law Variable

<table>
<thead>
<tr>
<th>Disability Law in Effect</th>
<th>Number</th>
<th>Uninsured</th>
<th>Non Group</th>
<th>Government</th>
<th>Group</th>
<th>Odds - comparing Group to Uninsured</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>82,246</td>
<td>20.1</td>
<td>12.2</td>
<td>8.4</td>
<td>59.3</td>
<td>2.946</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4,229</td>
<td>19.9</td>
<td>10.1</td>
<td>12.5</td>
<td>57.6</td>
<td>2.895</td>
<td>0.983</td>
</tr>
<tr>
<td>Total</td>
<td>86,475</td>
<td>20.1</td>
<td>12.1</td>
<td>8.6</td>
<td>59.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In contrast, Table 11.8 suggests disability laws may have important effects on the average health insurance purchasing behavior of selected subgroups of the sample. Table 11.8 shows the percent uninsured and odds of purchasing group insurance (compared to being uninsured) for selected subgroups. To illustrate, for disabled individuals, the odds of purchasing group insurance are 1.329 times higher when disability restrictions are in effect. Table 11.7 suggests that disability restrictions have no effect; this may be true when looking at the entire sample. However, by examining subgroups, Table 11.8 shows that we may see important effects associated with legal underwriting restrictions that are not evident when looking at averages over the whole sample.

Table 11.8. Odds of Health Coverage by Law and Physical Impairment

<table>
<thead>
<tr>
<th>Selected Subgroups</th>
<th>Disability Law in Effect</th>
<th>Number</th>
<th>Percent Group</th>
<th>Percent Uninsured</th>
<th>Odds - comparing Group to Uninsured</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nondisabled</td>
<td>No</td>
<td>72,150</td>
<td>64.2</td>
<td>20.5</td>
<td>3.134</td>
<td></td>
</tr>
<tr>
<td>Nondisabled</td>
<td>Yes</td>
<td>3,649</td>
<td>63.4</td>
<td>21.2</td>
<td>2.985</td>
<td>0.952</td>
</tr>
<tr>
<td>Disabled</td>
<td>No</td>
<td>10,096</td>
<td>24.5</td>
<td>17.6</td>
<td>1.391</td>
<td></td>
</tr>
<tr>
<td>Disabled</td>
<td>Yes</td>
<td>580</td>
<td>21.0</td>
<td>11.4</td>
<td>1.848</td>
<td>1.329</td>
</tr>
</tbody>
</table>

There are many ways of picking subgroups of interest. With a large dataset of $n = 86,475$ observations, one could probably pick subgroups to confirm almost any hypothesis. Further, there is a concern that the CPS data may not provide a
nominal dependent variables

representative sample of state populations. Thus, it is customary to use regression
techniques to “control” for explanatory variables, such as physical impairment.

Table 11.9 reports the main results from a multinomial logit model with many
control variables included. A dummy variable for each of 50 states was included
(the District of Columbia is a “state” in this data set, so we need 51 − 1 = 50
dummy variables). These variables were suggested in the literature and are further
described in Browne and Frees (2007). They include an individual’s gender, marital
status, race, education, whether or not self-employed and whether an individual
worked full-time, part-time or not at all.

In Table 11.9, “Law” refers to the binary variable that is 1 if a legal restriction was
in effect and “Disabled” is a binary variable that is 1 if an individual is physically
impaired. Thus, the interaction “Law*Disabled” reports the effect of a legal restric-
tion on a physically impaired individual. The interpretation is similar to Table 11.8.
Specifically, we interpret the coefficient 1.419 to mean that disabled individuals are
41.9% more likely to purchase group health insurance compared to purchasing no
insurance, when the disability underwriting restriction is in effect. Similarly, non-
disabled individuals are 21.2% (= 1/0.825 − 1) less likely to purchase group health
insurance compared to purchasing no insurance, when the disability underwriting
restriction is in effect. This result suggests that the non-disabled are more likely to
be uninsured as a result of prohibitions on the use of disability status as an under-
writing criteria. Overall, the results are statistically significant, confirming that this
legal restriction does have an impact on the consumption of health insurance.

Table 11.9. Odds Ratios from Multinomial Logit Regression Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group versus Uninsured</th>
<th>Group versus Non-Group Uninsured</th>
<th>Group versus Government Uninsured</th>
<th>Group versus Non-Group Government Uninsured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law*Nondisabled</td>
<td>0.825</td>
<td>1.053</td>
<td>1.010</td>
<td>0.784</td>
</tr>
<tr>
<td>p-value</td>
<td>0.001</td>
<td>0.452</td>
<td>0.900</td>
<td>0.001</td>
</tr>
<tr>
<td>Law*Disabled</td>
<td>1.419</td>
<td>0.953</td>
<td>1.664</td>
<td>1.490</td>
</tr>
<tr>
<td>p-value</td>
<td>0.062</td>
<td>0.789</td>
<td>0.001</td>
<td>0.079</td>
</tr>
<tr>
<td>Notes: The regression includes 150 (= 50 × 3) state-specific effects, several continuous variables (age, education and income, as well as higher order terms) and categorical variables (such as race and year).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11.5.3 Nested Logit

To mitigate the problem of independence of irrelevant alternatives in multinomial
logits, we now introduce a type of hierarchical model known as a nested logit model.
To interpret the nested logit model, in the first stage one chooses an alternative (say
the first alternative) with probability
\[
\pi_{i,1} = \Pr(y_i = 1) = \frac{\exp(V_{i,1})}{\exp(V_{i,1}) + \sum_{k=2}^{c} \exp(V_{i,k}/\rho)).
\]
Then, conditional on not choosing the first alternative, the probability of choosing
any one of the other alternatives follows a multinomial logit model with probabilities
\[
\frac{\pi_{i,j}}{1 - \pi_{i,1}} = \Pr(y_i = j | y_i \neq 1) = \frac{\exp(V_{i,j}/\rho)}{\sum_{k=2}^{c} \exp(V_{i,k}/\rho)), \quad j = 2, \ldots, c.
\]
22  

Categorical Dependent Variables

In equations (11.9) and (11.10), the parameter $\rho$ measures the association among the choices $j = 2, \ldots, c$. The value of $\rho = 1$ reduces to the multinomial logit model that we interpret to mean independence of irrelevant alternatives. We also interpret Prob($y_i = 1$) to be a weighted average of values from the first choice and the others. Conditional on not choosing the first category, the form of Pr($y_i = j | y_i \neq 1$) in equation (11.10) has the same form as the multinomial logit.

The advantage of the nested logit is that it generalizes the multinomial logit model in a way such that we no longer have the problem of independence of irrelevant alternatives. A disadvantage, pointed out by McFadden (1981), is that only one choice is observed; thus, we do not know which category belongs in the first stage of the nesting without additional theory regarding choice behavior. Nonetheless, the nested logit generalizes the multinomial logit by allowing alternative “dependence” structures. That is, one may view the nested logit as a robust alternative to the multinomial logit and examine each one of the categories in the first stage of the nesting.

11.6 Ordinal Dependent Variables

We now consider a response that is an ordered categorical variable, also known as an ordinal dependent variable. To illustrate, any type of survey response where you score your impression on a seven point scale ranging from “very dissatisfied” to “very satisfied” is an example of an ordinal variable.

Example: Health Plan Choice. Pauly and Herring (2007) examined $c = 4$ choices of health care plan types, consisting of:

- $y = 1$ - a health maintenance organization (HMO),
- $y = 2$ - a point of service (POS) plan,
- $y = 3$ - a preferred provider organization (PPO) or
- $y = 4$ - a fee for service (FFS) plan.

A FFS plan is the least restrictive, allowing enrollees to see health care providers (such as primary care physicians) for a fee reflecting the cost of services rendered. The PPO plan is the next least restrictive; this plan generally uses FFS payments but enrollees generally must choose from a list of “preferred providers.” Pauly and Herring (2007) took POS and HMO plans to be the third and fourth least restrictive, respectively. An HMO often uses capitation (a flat rate per person) to reimburse providers, restricting enrollees to a network of providers. In contrast, a POS plan gives enrollees the option to see providers outside of the HMO network (for an additional fee).

11.6.1 Cumulative Logit

Models of ordinal dependent variables are based on cumulative probabilities of the form

$$\Pr(y \leq j) = \pi_1 + \cdots + \pi_j, \quad j = 1, \ldots, c.$$
In this section, we use cumulative logits

\[
\text{logit} \left( \Pr(y \leq j) \right) = \ln \left( \frac{\Pr(y \leq j)}{1 - \Pr(y \leq j)} \right) = \ln \left( \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c} \right). \tag{11.11}
\]

The simplest cumulative logit model is

\[
\text{logit} \left( \Pr(y \leq j) \right) = \alpha_j
\]

that does not use any explanatory variables. The “cut-point” parameters \( \alpha_j \) are nondecreasing so that \( \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_c \), reflecting the cumulative nature of the distribution function \( \Pr(y \leq j) \).

The proportional odds model incorporates explanatory variables. With this model, cumulative logits are expressed as

\[
\text{logit} \left( \Pr(y \leq j) \right) = \alpha_j + x_i' \beta. \tag{11.12}
\]

This model provides parameter interpretations similar to those for logistic regression described in Section 11.2.4. For example, if the variable \( x_1 \) is continuous, then as in equation (11.1) we have

\[
\beta_1 = \frac{\partial}{\partial x_{1i}} (\alpha_j + x_i' \beta) = \frac{\partial}{\partial x_{1i}} \frac{\Pr(y_i \leq j \mid x_i) / (1 - \Pr(y_i \leq j \mid x_i))}{\Pr(y_i \leq j \mid x_i) / (1 - \Pr(y_i \leq j \mid x_i))}.
\]

Thus, we may interpret \( \beta_1 \) as the proportional change in the cumulative odds ratio.

**Example: Health Plan Choice - Continued.** Pauly and Herring used data from the 1996-1997 and 1998-1999 Community Tracking Study’s Household Surveys (CTS-HS) to study the demand for health insurance. This is a nationally representative survey containing over 60,000 individuals per period. As one measure of demand, Pauly and Herring examined health plan choice, reasoning that individuals that chose (through employment or association membership) less restrictive plans sought greater protection for health care. (They also looked at other measures, including the number of restrictions placed on plans and the amount of cost-sharing.) Table 11.10 provides determinants of health plan choice based on \( n = 34,486 \) individuals who had group health insurance, aged 18-64 without public insurance. Pauly and Herring also compared these results to those who had individual health insurance to understand the differences in determinants between these two markets.

| Table 11.10. Cumulative Logit Model of Health Plan Choice |
|-----------------|-----------------|-----------------|-----------------|
| Variable        | Odds Ratio      | Variable        | Odds Ratio      |
| Age             | 0.992***        | Hispanic        | 1.735***        |
| Female          | 1.064***        | Risk taker      | 0.967           |
| Family size     | 0.985           | Smoker          | 1.055***        |
| Family income   | 0.963***        | Fair/poor health| 1.056           |
| Education       | 1.006           | \( \alpha_1 \)  | 0.769***        |
| Asian           | 1.180***        | \( \alpha_2 \)  | 1.406***        |
| African-American| 1.643***        | \( \alpha_3 \)  | 12.089***       |
| Maximum-rescaled \( R^2 \) | 0.102 |

Notes: Source: Pauly and Herring (2007). *** indicates that the associated \( p \)-values are less than 0.01. For race, Caucasian is the omitted variable.
24

Categorical Dependent Variables

To interpret the odds ratios in Table 11.10, we first note that the cut-point estimates, corresponding to $\alpha_1$, $\alpha_2$ and $\alpha_3$, increase as choices become less restrictive, as anticipated. For gender, we see that the estimated odds for females are 1.064 times that of males in the direction of choosing a less restrictive health plan. Controlling for other variables, females are more likely to choose less restrictive plans than males. Similarly, younger, less wealthy, non-Caucasian and smokers are more likely to choose less restrictive plans. Coefficients associated with family size, education, risk taking and self reported health were not statistically significant in this fitted model.

11.6.2 Cumulative Probit

As in Section 11.2.2 for logistic regression, cumulative logit models have a threshold interpretation. Specifically, let $y_i^*$ be a latent, unobserved, random variable upon which we base the observed dependent variable as

$$y_i = \begin{cases} 1 & y_i^* \leq \alpha_1 \\ 2 & \alpha_1 < y_i^* \leq \alpha_2 \\ \vdots & \vdots \\ c-1 & \alpha_{c-2} < y_i^* \leq \alpha_{c-1} \\ c & \alpha_{c-1} < y_i^* \end{cases}$$

If $y_i^* - x_i'\beta$ has a logistic distribution, then

$$\Pr(y_i^* - x_i'\beta \leq a) = \frac{1}{1 + \exp(-a)}$$

and thus

$$\Pr(y_i \leq j) = \Pr(y_i^* \leq \alpha_j) = \frac{1}{1 + \exp(-\alpha_j - x_i'\beta)}.$$ 

Applying the logit transform to both sides yields equation (11.12).

Alternatively, assume that $y_i^* - x_i'\beta$ has a standard normal distribution. Then,

$$\Pr(y_i \leq j) = \Pr(y_i^* \leq \alpha_j) = \Phi(\alpha_j - x_i'\beta).$$

This is the cumulative probit model. As with binary variable models, the cumulative probit gives results that are similar to the cumulative logit model.

11.7 Further Reading and References

Regression models of binary variables are used extensively. For more detailed introductions, see Hosmer and Lemshow (1989) or Agresti (1996). You may also wish to examine more rigorous treatments such as those in Agresti (1990) and Cameron and Trivedi (1998). The work by Agresti (1990, 1996) discuss multicategory dependent variables, as does the advanced econometrics treatment in Amemiya (1985).

Chapter References
11.8 Exercises

Exercises

11.1 Similarity of Logit and Probit. Suppose that the random variable $y^*$ has a logit distribution function, $\Pr(y^* \leq y) = F(y) = e^y/(1 + e^y)$.

a. Calculate the corresponding probability density function.

b. Use the probability density function to compute the mean ($\mu_y$).

c. Compute the corresponding standard deviation ($\sigma_y$).

d. Define the rescaled random variable $y^{**} = y^* - \mu_y/\sigma_y$. Determine the probability density function for $y^{**}$.

e. Plot the probability density function in part (d). Overlay this plot with a plot of a standard normal probability density function. (This provides a density function version of the distribution function plots in Figure 11.1.)

11.2 Threshold interpretation of the probit regression model. Consider an underlying linear model, $y^*_i = x'_i \beta + \epsilon^*_i$, where $\epsilon^*_i$ is normally distributed with mean zero and variance $\sigma^2$. Define $y_i = I(y^*_i > 0)$, where $I(\cdot)$ is the indicator function. Show that $\pi_i = \Pr(y_i = 1|x_i) = \Phi(x'_i \beta/\sigma)$, where $\Phi(\cdot)$ is the standard normal distribution function.

11.3 Random utility interpretation of the logistic regression model. Under the random utility interpretation, an individual with utility $U_{ij} = u_i(V_{ij} + \epsilon_{ij})$, where $j$ may be 1 or 2, selects category corresponding to $j = 1$ with probability

$$
\pi_i = \Pr(y_i = 1) = \Pr(U_{i2} < U_{i1}) = \Pr(\epsilon_{i2} - \epsilon_{i1} < V_{i1} - V_{i2}).
$$

As in Section 11.2.3, we take $V_{i2} = 0$ and $V_{i1} = x'_i \beta$. Further suppose that the errors are from an extreme value distribution of the form

$$
\Pr(\epsilon_{ij} < a) = \exp(-e^{-a}).
$$
Show that the choice probability $\pi_i$ has a logit form. That is, show

$$\pi_i = \frac{1}{1 + \exp(-x'_i \beta)}.$$ 

11.4 Two Populations.

a. Begin with one population and assume that $y_1, \ldots, y_n$ is an i.i.d. sample from a Bernoulli distribution with mean $\pi$. Show that the maximum likelihood estimator of $\pi$ is $\bar{y}$.

b. Now consider two populations. Suppose that $y_1, \ldots, y_{n_1}$ is an i.i.d. sample from a Bernoulli distribution with mean $\pi_1$ and $y_{n_1+1}, \ldots, y_{n_1+n_2}$ is an i.i.d. sample from a Bernoulli distribution with mean $\pi_2$, where the samples are independent of one another.

b(i). Show that the maximum likelihood estimator of $\pi_2 - \pi_1$ is $\bar{y}_2 - \bar{y}_1$.

b(ii). Determine the variance of the estimator in part b(i).

c. Now express the two population problem in a regression context using one explanatory variable. Specifically, suppose that $x_i$ only takes on the values 0 and 1. Out of the $n$ observations, $n_1$ take on the value $x = 0$. These $n_1$ observations have an average $y$ value of $\bar{y}_1$. The remaining $n_2 = n - n_1$ observations have value $x = 1$ and an average $y$ value of $\bar{y}_2$. Using the logit case, let $b_{0,\text{MLE}}$ and $b_{1,\text{MLE}}$ represent the maximum likelihood estimators of $\beta_0$ and $\beta_1$, respectively.

c(i). Show that the maximum likelihood estimators satisfy the equations

$$\bar{y}_1 = \pi(b_{0,\text{MLE}})$$

and

$$\bar{y}_2 = \pi(b_{0,\text{MLE}} + b_{1,\text{MLE}}).$$

c(ii). Use part c(i) to show that the maximum likelihood estimator for $\beta_1$ is $\pi^{-1}(\bar{y}_2) - \pi^{-1}(\bar{y}_1)$.

c(iii). With the notation $\pi_1 = \pi(\beta_0)$ and $\pi_2 = \pi(\beta_0 + \beta_1)$, confirm that the information matrix can be expressed as

$$I(\beta_0, \beta_1) = n_1 \pi_1 (1 - \pi_1) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + n_2 \pi_2 (1 - \pi_2) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$ 

c(iv). Use the information matrix to determine the large sample variance of the maximum likelihood estimator for $\beta_1$.

11.5 Fitted Values. Let $\hat{y}_i = \pi(x'_i b_{\text{MLE}})$ denote the $i$th fitted value for the logit function. Assume that an intercept is used in the model so that one of the explanatory variables $x$ is a constant equal to one. Show that the average response is equal to the average fitted value, that is, show $\bar{y} = n^{-1} \sum_{i=1}^n \hat{y}_i$.

11.6 Beginning with the score equations (11.4), verify the expression for the logit case in equation (11.5).

11.7 Information Matrix

a. Beginning with the score function for the logit case in equation (11.5), show that the information matrix can be expressed as

$$I(\beta) = \sum_{i=1}^n \sigma_i^2 x'_i x'_i,$$

where $\sigma_i^2 = \pi(x'_i \beta)(1 - \pi(x'_i \beta))$.

b. Beginning with the general score function in equation (11.4), determine the information matrix.
11.8 Automobile injury insurance claims. Refer to the description in Exercise 1.7.??.

We consider \( n = 1,340 \) bodily injury liability claims from a single state using a 2002 survey conducted by the Insurance Research Council (IRC). The IRC is a division of the American Institute for Chartered Property Casualty Underwriters and the Insurance Institute of America. The survey asked participating companies to report claims closed with payment during a designated two week period. In this assignment, we are interested in understanding the characteristics of the claimants who choose to be presented by an attorney when settling their claim. Variable descriptions are given Table 11.11.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTORNEY</td>
<td>whether the claimant is represented by an attorney (=1 if yes and =2 if no)</td>
</tr>
<tr>
<td>CLMAGE</td>
<td>claimant’s age</td>
</tr>
<tr>
<td>CLMSEX</td>
<td>claimant’s gender (=1 if male and =2 if female)</td>
</tr>
<tr>
<td>MARITAL</td>
<td>claimant’s marital status (=1 if married, =2 if single, =3 if widowed, and =4 if divorced/separated)</td>
</tr>
<tr>
<td>SEATBELT</td>
<td>whether or not the claimant was wearing a seatbelt/child restraint (=1 if yes, =2 if no, and =3 if not applicable)</td>
</tr>
<tr>
<td>CLMINSUR</td>
<td>whether or not the driver of the claimant’s vehicle was uninsured (=1 if yes, =2 if no, and =3 if not applicable)</td>
</tr>
<tr>
<td>LOSS</td>
<td>the claimant’s total economic loss (in thousands).</td>
</tr>
</tbody>
</table>

a. **Summary Statistics.**

i. Calculate histograms and summary statistics of continuous explanatory variables CLMAGE and LOSS. Based on these results, create a logarithm version of LOSS, say lnLOSS.

ii. Examine the means of CLMAGE, LOSS and lnLOSS by level of ATTORNEY. Do these statistics suggest that the continuous variables differ by ATTORNEY?

iii. Create tables of counts (or percentages) of ATTORNEY by level of CLMSEX, MARITAL, SEATBELT, and CLMINSUR. Do these statistics suggest that the categorial variables differ by ATTORNEY?

iv. Identify the number of missing values for each explanatory variable.

b. **Logistic Regression Models.**

i. Run a logistic regression model using only the explanatory variable CLMSEX. Is it an important factor in determining the use of an attorney? Provide an interpretation in terms of the odds of using an attorney.

ii. Run a logistic regression model using the explanatory variables CLMAGE, CLMSEX, MARITAL, SEATBELT, and CLMINSUR. Which variables appear to be statistically significant?

iii. For the model in part (ii), who uses attorneys more, men or women? Provide an interpretation in terms of the odds of using an attorney for the variable CLMSEX.

iv. Run a logistic regression model using the explanatory variables CLMAGE, CLMSEX, MARITAL, SEATBELT, CLMINSUR, LOSS and lnLOSS. Decide which of the two loss measures is more important and re-run the model using only one of these variables. In this model, is the measure of losses a statistically significant variable?

v. Run your model in part (iv) but omitting the variable CLMAGE. Describe differences between this model fit and that in part (iv), focusing on statistically significant variables and number of observations used in the model fit.

vi. Consider a single male claimant who is age 32. Assume that the claimant was wearing a seat belt, that the driver was insured and the total economic loss is...
For the model in part (iv), what is the estimate of the probability of using an attorney?

c. Probit Regression. Repeat part b(v) using probit regression models but interpret only the sign of the regression coefficients.

## 11.9 Hong Kong Horse Racing

The race track is a fascinating example of financial market dynamics at work. Let’s go to the track and make a wager. Suppose that, from a field of 10 horses, we simply want to pick a winner. In the context of regression, we will let \( y \) be the response variable indicating whether a horse wins \( (y = 1) \) or not \( (y = 0) \). From racing forms, newspapers and so on, there are many explanatory variables that are publicly available that might help us predict the outcome for \( y \). Some candidate variables may include the age of the horse, recent track performance of the horse and jockey, pedigree of the horse, and so on. These variables are assessed by the investors present at the race, the betting crowd. Like many financial markets, it turns out that one of the most useful explanatory variable is the crowd’s overall assessment of the horse’s abilities. These assessments are not made based on a survey of the crowd, but rather based on the wagers placed. Information about the crowd’s wagers is available on a large sign at the race called the **tote board**. The tote board provides the odds of each horse winning a race. Table 11.12 is a hypothetical tote board for a race of 10 horses.

<table>
<thead>
<tr>
<th>Horse</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posted Odds</td>
<td>1-1</td>
<td>79-1</td>
<td>7-1</td>
<td>3-1</td>
<td>15-1</td>
<td>7-1</td>
<td>49-1</td>
<td>49-1</td>
<td>19-1</td>
<td>79-1</td>
</tr>
</tbody>
</table>

The odds that appear on the tote board have been adjusted to provide a “track take.” That is, for every dollar that has been wagered, \( $T \) goes to the track for sponsoring the race and \( $(1-T) \) goes to the winning bettors. Typical track takes are in the neighborhood of twenty percent, or \( T=0.20 \).

We can readily convert the odds on the tote board to the crowd’s assessment of the probabilities of winning. To illustrate this, Table 11.13 shows hypothetical bets to win which resulted in the displayed information on the hypothetical tote board in Table 11.12.

<table>
<thead>
<tr>
<th>Horse</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bets to Win</td>
<td>8,000</td>
<td>200</td>
<td>2,000</td>
<td>4,000</td>
<td>1,000</td>
<td>3,000</td>
<td>400</td>
<td>400</td>
<td>800</td>
<td>200</td>
<td>20,000</td>
</tr>
<tr>
<td>Probability</td>
<td>0.40</td>
<td>0.01</td>
<td>0.10</td>
<td>0.20</td>
<td>0.05</td>
<td>0.15</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>1.000</td>
</tr>
<tr>
<td>Posted Odds</td>
<td>1-1</td>
<td>79-1</td>
<td>7-1</td>
<td>3-1</td>
<td>15-1</td>
<td>7-1</td>
<td>49-1</td>
<td>49-1</td>
<td>19-1</td>
<td>79-1</td>
<td></td>
</tr>
</tbody>
</table>

For this hypothetical race, $20,000 was bet to win. Because $8,000 of this $20,000 was bet on the first horse, interpret the ratio \( 8000/20000 = 0.40 \) as the crowd’s assessment of the probability to win. The theoretical odds are calculated as \( 0.4/(1-0.4) = 2/3 \), or a 0.67 bet wins $1. However, the theoretical odds assume a fair game with no track take. To adjust for the fact that only \( $(1-T) \) are available to the winner, the posted odds for this horse would be \( 0.4/(1-T-0.4) = 1 \), if \( T=0.20 \). For this case, it now takes a $1 bet to win $1. We then have the relationship adjusted odds \( = x/(1-T-x) \), where \( x \) is the crowd’s assessment of the probability of winning.

Before the start of the race, the tote board provides us with adjusted odds that
Exercises

29

can readily be converted into $x$, the crowd’s assessment of winning. We use this measure to help us to predict $y$, the event of the horse actually winning the race.

We consider data from 925 races run in Hong Kong from September, 1981 through September, 1989. In each race, there were ten horses, one of whom was randomly selected to be in the sample. In the data, use $\text{FINISH} = y$ to be the indicator of a horse winning a race and $\text{WIN} = x$ to be the crowd’s a priori probability assessment of a horse winning a race.

a. A statistically naive colleague would like to double the sample size by picking two horses from each race instead of randomly selecting one horse from a field of 10.

i. Describe the relationship between the dependent variables of the two horses selected.

ii. Say how this violates the regression model assumptions.

b. Calculate the average $\text{FINISH}$ and summary statistics for $\text{WIN}$. Note that the standard deviation of $\text{FINISH}$ is higher than that of $\text{WIN}$, even though the sample means are about the same. For the variable $\text{FINISH}$, what is the relationship between the sample mean and standard deviation?

c. Calculate summary statistics of $\text{WIN}$ by level of $\text{FINISH}$. Note that the sample mean is larger for horses that won ($\text{FINISH} = 1$) than for those that lost ($\text{FINISH} = 0$). Interpret this result.

d. Estimate a linear probability model, using $\text{WIN}$ to predict $\text{FINISH}$.

i. Is $\text{WIN}$ a statistically significant predictor of $\text{FINISH}$?

ii. How well does this model fit the data using the usual goodness of fit statistic?

iii. For this estimated model, is it possible for the fitted values to lie outside the interval $[0, 1]$? Note, by definition, that the x-variable $\text{WIN}$ must lie within the interval $[0, 1]$.

e. Estimate a logistic regression model, using $\text{WIN}$ to predict $\text{FINISH}$. Is $\text{WIN}$ a statistically significant predictor of $\text{FINISH}$?

f. Compare the fitted values from the models in parts (d) and (e)

i. For each model, provide fitted values at $\text{WIN} = 0, 0.01, 0.05, 0.10$ and 1.0.

ii. Plot fitted values from the linear probability model versus fitted values from the logistic regression model.

g. Interpret $\text{WIN}$ as the crowd’s prior probability assessment of the probability of a horse winning a race. The fitted values, $\text{FINISH}$, is your new estimate of the probability of a horse winning a race, based on the crowd’s assessment.

i. Plot the difference $\text{FINISH} - \text{WIN}$ versus $\text{WIN}$.

ii. Discuss a betting strategy that you might employ based on the difference, $\text{FINISH} - \text{WIN}$.

11.10 Demand for Term Life Insurance. We continue our study of Term Life Insurance Demand from Chapters 3 and 4. Specifically, we examine the 2004 Survey of Consumer Finances (SCF), a nationally representative sample that contains extensive information on assets, liabilities, income, and demographic characteristics of those sampled (potential U.S. customers). We now return to the original sample of $n = 500$ families with positive incomes and study whether or not a family purchases term life insurance. From our sample, it turns out that 225 did not (FACEPOS=0), whereas 275 did purchase term life insurance (FACEPOS=1).

a. Summary Statistics. Provide a table of means of explanatory variables by level of the dependent variable FACEPOS. Interpret what we learn from this table.

b. Linear Probability Model. Fit a linear probability model using FACEPOS as the dependent variable and LINCOME, EDUCATION, AGE and GENDER as continuous explanatory variables, together with the factor MARSTAT.

b(i). Briefly define a linear probability model.

b(ii). Comment on the quality of the fitted model.
b(iii). What are the three main drawbacks of the linear probability model?
c. Logistic Regression Model. Fit a logistic regression model using the same set of explanatory variables.
c(i). Identify which variables appear to be statistically significant. In your identification, describe the basis for your conclusions.
c(ii). Which measure summarizes the goodness of fit?
d. Reduced Logistic Regression Model. Define MARSTAT1 to be a binary variable that indicates MARSTAT=1. Fit a second logistic regression model using LINCOME, EDUCATION and MARSTAT1.
d(i). Compare these two models, using a likelihood ratio test. State your null and alternative hypotheses, decision making criterion and your decision-making rule.
d(ii). Who is more likely to purchase term life insurance, married or “non” married? Provide an interpretation in terms of the odds of purchasing term life insurance for the variable MARSTAT1.
d(iii). Consider a married male who is age 54. Assume that this person has 13 years of education, annual wages of $70,000 and is living in a household composed of four people. For this model, what is the estimate of the probability of purchasing term life insurance?

11.11 Success in Actuarial Studies. Much like the medical and legal fields, members of the actuarial profession face interesting problems and are generally well compensated for their efforts in resolving these problems. Also like the medical and legal professions, the educational barriers to becoming an actuary are challenging, limiting entrance into the field.

To advise students on whether they have the potential to meet the demands of this intellectually challenging field, Smith and Schumacher (2006) studied attributes of students in a business college. Specifically, they examined \( n = 185 \) freshman at Bryant University in Rhode Island who had begun their college careers in 1995-2001. The dependent variable of interest was whether they graduated with an actuarial concentration, for these students the first step to becoming a professional actuary. Of these, 77 graduated with an actuarial concentration and the other 108 dropped the concentration (at Bryant, most transferred to other concentrations although some left the university).

Smith and Schumacher (2006) reported the effects of four early assessment mechanisms as well as GENDER, a control variable. The assessment mechanisms were: PLACE%, performance on a mathematics placement exam administered just prior to the freshman year, MSAT and VSAT, mathematics (M) and verbal (V) portions of the Scholastic Aptitude Test (SAT) and RANK, high school rank given as a proportion (with closer to one being better). Table 11.14 shows that students who eventually graduated with an actuarial concentration and the other 108 dropped the concentration (at Bryant, most transferred to other concentrations although some left the university).

A logistic regression was fit to the data with the results reported in Table 11.14.
a. To get a sense of which variables are statistically significant, calculate \( t \)-ratios for each variable. For each variable, state whether or not it is statistically significant.
b. To get a sense of the relative impact of the assessment mechanisms, use the coefficients in Table 11.14 to compute estimated success probabilities for the following combination of variables. In your calculations, assume that GENDER = 1.
b(i). Assume PLACE% = 0.80, MSAT = 680, VSAT = 570 and RANK = 0.90.
b(ii). Assume PLACE% = 0.60, MSAT = 680, VSAT = 570 and RANK = 0.90.
b(iii). Assume PLACE% = 0.80, MSAT = 620, VSAT = 570 and RANK = 0.90.
b(iv). Assume PLACE% = 0.80, MSAT = 680, VSAT = 540 and RANK = 0.90.
b(v). Assume PLACE% = 0.80, MSAT = 680, VSAT = 570 and RANK = 0.70.
Case-Control. Consider the following “case-control” sample selection method for binary dependent variables. Intuitively, if we are working with a problem where the event of interest is rare, we want to make sure that we sample a sufficient number of events so that our estimation procedures are reliable.

Suppose that we have a large database consisting of \( \{y_i, x_i\}, i = 1, \ldots, N \) observations. (For insurance company records, \( N \) could easily be ten million or more.) We want to make sure to get plenty of \( y_i = 1 \) (corresponding to claims or “cases”) in our sample, plus a sample of \( y_i = 0 \) (corresponding to non-claims or “controls”). Thus, we split the data set into two subsets. For the first subset consisting of observations with \( y_i = 1 \), we take a random sample with probability \( \tau_1 \). Similarly, for the second subset consisting of observations with \( y_i = 0 \), we take a random sample with probability \( \tau_0 \). For example, in practice we might use \( \tau_1 = 1 \) and \( \tau_0 = 0.1 \), corresponding to taking all of the claims and a 10% sample of non-claims - thus, \( \tau_1 \) and \( \tau_1 \) are considered known to the analyst.

a. Let \( \{r_i = 1\} \) denote the event that the observation is selected to be part of the analysis. Determine \( \Pr(y_i = 1, r_i = 1) \), \( \Pr(y_i = 0, r_i = 1) \) and \( \Pr(r_i = 1) \) in terms of \( \tau_0 \), \( \tau_1 \) and \( \pi_i = \Pr(y_i = 1) \).

b. Using the calculations in part (a), determine the conditional probability \( \Pr(y_i = 1| r_i = 1) \).

c. Now assume that \( \pi_i \) has a logistic form (\( \pi(z) = \exp(z)/(1 + \exp(z)) \)) and \( \pi_i = \pi(x_i' \beta) \). Re-write your answer part (b) using this logistic form.

d. Write the likelihood of the observed \( y_i \)'s (conditional on \( r_i = 1, i = 1, \ldots, n \)). Show how we can interpret this as the usual logistic regression likelihood with the exception that the intercept has changed. Specify the new intercept in terms of the original intercept, \( \tau_0 \) and \( \tau_1 \).

11.9 Technical Supplements - Likelihood-Based Inference

Begin with random variables \( (y_1, \ldots, y_n)' = y \) whose joint distribution is known up to a vector of parameters \( \theta \). In regression applications, \( \theta \) consists of the regression coefficients, \( \beta \), and possibly a scale parameter \( \sigma^2 \) as well as additional parameters. This joint probability density function is denoted as \( f(y; \theta) \). The function may also be a probability mass function for discrete random variables or a mixture distribution for random variables that have discrete and continuous components. In each case, we can use the same notation, \( f(y; \theta) \), and call it the likelihood function. The likelihood is a function of the parameters with the data \( (y) \) fixed rather than a function of the data with the parameters \( (\theta) \) fixed.

It is customary to work with the logarithmic version of the likelihood function.
and thus we define the log-likelihood function to be

\[ L(\theta) = L(y; \theta) = \ln f(y; \theta) , \]

evaluated at a realization of \( y \). In part, this is because we often work with the important special case where the random variables \( y_1, \ldots, y_n \) are independent. In this case, the joint density function can be expressed as a product of the marginal density functions and, by taking logarithms, we can work with sums. Even when not dealing with independent random variables, as with time series data, it is often computationally more convenient to work with log-likelihoods than the original likelihood function.

### 11.9.1 Properties of Likelihood Functions

Two basic properties of likelihood functions are:

\[ E \left( \frac{\partial}{\partial \theta} L(\theta) \right) = 0 \]

and

\[ E \left( \frac{\partial^2}{\partial \theta \partial \theta'} L(\theta) \right) + E \left( \frac{\partial L(\theta)}{\partial \theta} \frac{\partial L(\theta)}{\partial \theta'} \right) = 0. \]

The derivative of the log-likelihood function, \( \frac{\partial L(\theta)}{\partial \theta} \), is called the score function. Equation (11.1) shows that the score function has mean zero. To see this, under suitable regularity conditions, we have

\[ E \left( \frac{\partial}{\partial \theta} f(y; \theta) \right) = \int \frac{\partial}{\partial \theta} f(y; \theta) dy = \frac{\partial}{\partial \theta} \int f(y; \theta) dy = 0. \]

For convenience, this demonstration assumes a density for \( f(\cdot) \); extensions to mass and mixtures distributions are straightforward. The proof of equation (11.2) is similar and is omitted. To establish equation (11.1), we implicitly used “suitable regularity conditions” to allow the interchange of the derivative and integral sign. To be more precise, an analyst working with a specific type of distribution can use this information to check that the interchange of the derivative and integral sign is valid.

Using equation (11.2), we can define the information matrix

\[ I(\theta) = E \left( \frac{\partial L(\theta)}{\partial \theta} \frac{\partial L(\theta)}{\partial \theta'} \right) = -E \left( \frac{\partial^2}{\partial \theta \partial \theta'} L(\theta) \right). \]

This quantity is used extensively in the study of large sample properties of likelihood functions.

The information matrix appears in the large sample distribution of the score function. Specifically, under broad conditions, we have that \( \frac{\partial L(\theta)}{\partial \theta} \) has a large sample normal distribution with mean 0 and variance \( I(\theta) \). To illustrate, suppose that the random variables are independent so that the score function can be written as

\[ \frac{\partial}{\partial \theta} L(\theta) = \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(y_i; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(y_i; \theta). \]
The score function is the sum of mean zero random variables because of equation (11.1); central limit theorems are widely available to ensure that sums of independent random variables have large sample normal distributions (see Section 1.4 for an example). Further, if the random variables are identical, then from equation (11.3) we can see that the second moment of $\partial \ln f(y; \theta) / \partial \theta$ is the information matrix, yielding the result.

### 11.9.2 Maximum Likelihood Estimators

Maximum likelihood estimators are values of the parameters $\theta$ that are “most likely” to have been produced by the data. The value of $\theta$, say $\theta_{\text{MLE}}$, that maximizes $f(y; \theta)$ is called the maximum likelihood estimator. Because $\ln(\cdot)$ is a one-to-one function, we can also determine $\theta_{\text{MLE}}$ by maximizing the log-likelihood function, $L(\theta)$.

Under broad conditions, we have that $\theta_{\text{MLE}}$ has a large sample normal distribution with mean $\mu$ and variance $(I(\theta))^{-1}$. This is a critical result upon which much of estimation and hypothesis testing is based. To underscore this result, we examine the special case of “normal-based” regression.

**Special Case. Regression with normal distributions.** Suppose that $y_1, \ldots, y_n$ are independent and normally distributed, with mean $E y_i = \mu_i = x_i' \beta$ and variance $\sigma^2$. The parameters can be summarized as $\theta = (\beta', \sigma^2)'$. Recall from equation (1.1) that the normal probability density function is

$$f(y; \mu_i, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (y - \mu_i)^2 \right).$$

With this, the two components of the score function are

$$\frac{\partial}{\partial \beta} L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \ln f(y_i; x_i' \beta, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} (y_i - x_i' \beta)^2$$

and

$$\frac{\partial}{\partial \sigma^2} L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \sigma^2} \ln f(y_i; x_i' \beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - x_i' \beta)^2.$$ 

Setting these equations to zero and solving yields the maximum likelihood estimators

$$\beta_{\text{MLE}} = \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i y_i = b$$

and

$$\sigma^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' b)^2 = \frac{n - (k + 1)}{n} s^2.$$

Thus, the maximum likelihood estimator of $\beta$ is equal to the usual least squares estimator. The maximum likelihood estimator of $\sigma^2$ is a scalar multiple of the usual least squares estimator. The least squares estimators $s^2$ is unbiased whereas as $\sigma^2_{\text{MLE}}$ is only approximately unbiased in large samples.
The information matrix is
\[
I(\theta) = -E \left( \frac{\partial^2}{\partial \theta \partial \theta'} L(\theta) - \frac{\partial^2}{\partial \theta^2} L(\theta) \right) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i' x_i - 0 \right).
\]
Thus, \( \beta_{MLE} = b \) has a large sample normal distribution with mean \( \beta \) and variance-covariance matrix \( \sigma^2 (\sum_{i=1}^{n} x_i x_i')^{-1} \), as seen previously. Moreover, \( \sigma^2_{MLE} \) has a large sample normal distribution with mean \( \sigma^2 \) and variance \( 2\sigma^4/n \).

Maximum likelihood is a general estimation technique that can be applied in many statistical settings, not just regression and time series applications. It can be applied broadly and enjoys certain optimality properties. We have already cited the result that maximum likelihood estimators typically have a large sample normal distribution. Moreover, maximum likelihood estimators are the most efficient in the following sense. Suppose that \( \hat{\theta} \) is an alternative unbiased estimator. The Cramer-Rao theorem states, under mild regularity conditions, for all vectors \( c \), that
\[
\text{Var} c' \hat{\theta}_{MLE} \leq \text{Var} c' \hat{\theta}_{MLE}, \quad \text{for sufficiently large } n.
\]
We also note that \( 2 (L(\theta_{MLE}) - L(\theta)) \) has a chi-square distribution with degrees of freedom equal to the dimension of \( \theta \).

In a few applications, such as the regression case with a normal distribution, maximum likelihood estimators can be computed analytically as a closed-form expression. Typically, this can be done by finding roots of the first derivative of the function. However, in general, maximum likelihood estimators can not be calculated with closed form expressions and are determined iteratively. Two general procedures are widely used:

- **Newton-Raphson** uses the iterative algorithm
  \[
  \theta_{NEW} = \theta_{OLD} - \left( \frac{\partial^2 L}{\partial \theta \partial \theta'} \right)^{-1} \left( \frac{\partial L}{\partial \theta} \right) \Bigg|_{\theta = \theta_{OLD}}.
  \]  

- **Fisher scoring** uses the iterative algorithm
  \[
  \theta_{NEW} = \theta_{OLD} + I(\theta_{OLD})^{-1} \left( \frac{\partial L}{\partial \theta} \right) \Bigg|_{\theta = \theta_{OLD}}.
  \]

where \( I(\theta) \) is the information matrix.

**11.9.3 Hypothesis Tests**

We consider testing the null hypothesis \( H_0 : h(\theta) = d \), where \( d \) is a known vector of dimension \( r \times 1 \) and \( h(\cdot) \) is known and differentiable. This testing framework encompasses the general linear hypothesis introduced in Chapter 4 as a special case.

There are three general approaches for testing hypotheses, called the **likelihood ratio**, **Wald** and **Rao** tests. The Wald approach evaluates a function of the likelihood at \( \hat{\theta}_{MLE} \). The likelihood ratio approach uses \( \hat{\theta}_{MLE} \) and \( \hat{\theta}_{Reduced} \). Here, \( \hat{\theta}_{Reduced} \) is the value of \( \theta \) that maximizes \( L(\hat{\theta}_{Reduced}) \) under the constraint that \( h(\hat{\theta}) = d \). The Rao approach also uses \( \hat{\theta}_{Reduced} \) but determines it by maximizing \( L(\hat{\theta}) - \lambda' (h(\hat{\theta}) - d) \), where \( \lambda \) is a vector of Lagrange multipliers. Hence, Rao’s test is also called the **Lagrange multiplier test**.
The test statistics associated with the three approaches are:

- **LRT:** \( LRT = 2 \times \{ L(\theta_{MLE}) - L(\theta_{Reduced}) \} \)
- **Wald:** \( TS_W(\theta_{MLE}) \), where
  \[
  TS_W(\theta) = (h(\theta) - d)' \left\{ \frac{\partial}{\partial \theta} h(\theta)' (-I(\theta))^{-1} \frac{\partial}{\partial \theta} h(\theta) \right\}^{-1} (h(\theta) - d),
  \]
  and
- **Rao:** \( TS_R(\theta_{Reduced}) \), where
  \[
  TS_R(\theta) = \frac{\partial}{\partial \theta} L(\theta)' (-I(\theta))^{-1} \frac{\partial}{\partial \theta} L(\theta)'.
  \]

Under broad conditions, all three test statistics have large sample chi-square distributions with \( r \) degrees of freedom under \( H_0 \). All three methods work well when the number of parameters is finite dimensional and the null hypothesis specifies that \( \theta \) is on the interior of the parameter space.

The main advantage of the Wald statistic is that it only requires computation of \( \theta_{MLE} \) and not \( \theta_{Reduced} \). In contrast, the main advantage of the Rao statistic is that it only requires computation of \( \theta_{Reduced} \) and not \( \theta_{MLE} \). In many applications, computation of \( \theta_{MLE} \) is onerous. The likelihood ratio test is a direct extension of the partial \( F \)-test introduced in Chapter 4 - it allows one to directly compare nested models, a helpful technique in applications.

### 11.9.4 Information Criteria

Likelihood ratio tests are useful for choosing between two models that are nested, that is, where one model is a subset of the other. How do we compare models when they are not nested? One way is to use the following information criteria.

The distance between two probability distributions given by probability density functions \( g \) and \( f_{\theta} \) can be summarized by

\[
KL(g, f_{\theta}) = E_g \ln \frac{g(y)}{f_{\theta}(y)}.
\]

This is the Kullback-Leibler distance. Here, we have indexed \( f \) by a vector of parameters \( \theta \). If we let the density function \( g \) be fixed at a hypothesized value, say \( f_{\theta_0} \), then minimizing \( KL(f_{\theta_0}, f_{\theta}) \) is equivalent to maximizing the log-likelihood.

However, maximizing the likelihood does not impose sufficient structure on the problem because we know that we can always make the likelihood greater by introducing additional parameters. Thus, Akaike in 1974 showed that a reasonable alternative is to minimize

\[
AIC = -2 \times L(\theta_{MLE}) + 2 \times (\text{number of parameters}),
\]

known as Akaike’s Information Criterion. Here, the additional term \( 2 \times (\text{number of parameters}) \) is a penalty for the complexity of the model. With this penalty, one cannot improve upon the fit simply by introducing additional parameters. This statistic can be used when comparing several alternative models that are not necessarily nested. One picks the model that minimizes \( AIC \). If the models under consideration have the same number of parameters, this is equivalent to choosing the model that maximizes the log-likelihood.
We remark that this definition is not uniformly adopted in the literature. For example, in time series analysis, the \( AIC \) is rescaled by the number of parameters. Other versions that provide finite sample corrections are also available in the literature.

Schwarz in 1978 derived an alternative criterion using Bayesian methods. His measure is known as the *Bayesian Information Criterion*, defined as

\[
BIC = -2 \times \log(\theta_{MLE}) + (\text{number of parameters}) \times \ln(\text{number of observations}),
\]

This measure gives greater weight to the number of parameters. That is, other things being equal, \( BIC \) will suggest a more parsimonious model than \( AIC \).

Like the adjusted coefficient of determination \( R_a^2 \) that we have introduced in the regression literature, both \( AIC \) and \( BIC \) provide measures of fit with a penalty for model complexity. In normal linear regression models, Section 5.6 pointed out that minimizing \( AIC \) is equivalent to minimizing \( n \ln s^2 + k \). Another linear regression statistic that balances the goodness of fit and complexity of the model is Mallows \( C_p \) statistic. For \( p \) candidate variables in the model, this is defined as

\[
C_p = \frac{(\text{Error SS})_p}{s^2} - (n - 2p).
\]

See, for example, Cameron and Trivedi (1998) for references and further discussion of information criteria.
Frequency-Severity Models

Chapter Preview. Many data sets feature dependent variables that have a large proportion of zeros. This chapter introduces a standard econometric tool, known as a tobit model, for handling such data. The tobit model is based on observing a left-censored dependent variable, such as sales of a product or claim on a healthcare policy, where it is known that the dependent variable cannot be below zero. Although this standard tool can be useful, many actuarial data sets that feature a large proportion of zeros are better modeled in “two parts,” one part for the frequency and one part for the severity. This chapter introduces two-part models and provides extensions to an aggregate loss model, where a unit under study, such as an insurance policy, can result in more than one claim.

16.1 Introduction

Many actuarial data sets come in “two parts:”

- one part for the frequency, indicating whether or not a claim has occurred or, more generally, the number of claims and
- one part for the severity, indicating the amount of a claim.

In predicting or estimating claims distributions, we often associate the cost of claims with two components: the event of the claim and its amount, if the claim occurs. Actuaries term these the claims frequency and severity components, respectively. This is the traditional way of decomposing “two-part” data, where one can think of a zero as arising from a policy without a claim (Bowers et al., 1997, Chapter 2). Because of this decomposition, two-part models are also known as frequency-severity models. However, this formulation has been traditionally used without covariates to explain either the frequency or severity components. In the econometrics literature, Cragg (1971) introduced covariates into these two components, citing an example from fire insurance.

Healthcare data also often feature a large proportion of zeros that must be accounted for in the modeling. Zero values can represent an individual’s lack of healthcare utilization, no expenditure or non-participation in a program. In healthcare, Mullahy (1998) cites some prominent areas of potential applicability:

- outcomes research - amount of health care utilization or expenditures
- demand for health care - amount of health care sought, such as number of physician visits and
- substance abuse - amount consumed of tobacco, alcohol and illicit drugs.
The two-part aspect can be obscured by a natural way of recording data; enter the amount of the claim when the claim occurs (a positive number) and a zero for no claim. It is easy to overlook a large proportion of zeros, particularly when the analyst is also concerned with many covariates that may help explain a dependent variable. As we will see in this chapter, ignoring the two-part nature can lead to serious bias. To illustrate, recall from Chapter 6 a plot of individual’s income ($x$) versus amount of insurance purchased ($y$) (Figure 6.3). Fitting a single line to these data would misinform users about the effects of $x$ on $y$.

![Fig. 16.1. When individuals do not purchase insurance, they are recorded as $y = 0$ sales. The sample in this plot represents two subsamples, those who purchased insurance, corresponding to $y > 0$, and those who did not, corresponding to $y = 0$.](image)

In contrast, many insurers keep separate data files for frequency and severity. For example, insurers maintain a “policyholder” file that is established when a policy is underwritten. This file records much underwriting information about the insured(s), such as age, gender and prior claims experience, policy information such as coverage, deductibles and limitations, as well as the insurance claims event. A separate file, often known as the “claims” file, records details of the claim against the insurer, including the amount. (There may also be a “payments” file that records the timing of the payments although we shall not deal with that here.) This recording process makes it natural for insurers to model the frequency and severity as separate processes.

### 16.2 Tobit Model

One way of modeling a large proportion of zeros is to assume that the dependent variable is (left) censored at zero. This chapter introduces left-censored regression, beginning with the well-known *tobit model* that is based on the pioneering work of James Tobin (1958). Subsequently, Goldberger (1964) coined the phrase “tobit model,” acknowledging the work of Tobin and its similarity to the probit model.

As with probit (and other binary response) models, we use an unobserved, or latent, variable $y^*$ that is assumed to follow a linear regression model of the form

$$y^*_i = x'_i \beta + \epsilon_i. \quad (16.1)$$

A latent variable is not observed by the analyst.
The responses are censored or “limited” in the sense that we observe \( y_i = \max (y^*_i, d_i) \). The limiting value, \( d_i \), is a known amount. Many applications use \( d_i = 0 \), corresponding to zero sales or expenses, depending on the application. However, we also might use \( d_i \) for the daily expenses claimed for travel reimbursement and allow the reimbursement (such as $50 or $100) to vary by employee \( i \). Some readers may wish to review Section 14.2 for an introduction to censoring.

The model parameters consist of the regression coefficients, \( \beta \), and the variability term, \( \sigma^2 = \text{Var} \varepsilon_i \). With equation (16.1), we interpret the regression coefficients as the marginal change of \( E y^* \) per unit change in each explanatory variable. This may be satisfactory in some applications, such as when \( y^* \) represents an insurance loss.

However, for most applications, users are typically interested in marginal changes in \( E y \), that is, the expected value of the observed response. To interpret these marginal changes, it is customary to adopt the assumption of normality for the latent variable \( y^*_i \) (or equivalently for the disturbance \( \varepsilon_i \)). With this assumption, standard calculations (see Exercise 16.1) show that

\[
E y_i = d_i + \Phi \left( \frac{x'_i \beta - d_i}{\sigma} \right) (x'_i \beta - d_i + \sigma \lambda_i),
\]

where

\[
\lambda_i = \frac{\phi \left( (x'_i \beta - d_i)/\sigma \right)}{\Phi \left( (x'_i \beta - d_i)/\sigma \right)}.
\]

Here, \( \phi(.) \) and \( \Phi(.) \) are the standard normal density and distribution functions, respectively. The ratio of a probability density function to a cumulative distribution function is sometimes called an inverse Mills ratio. Although complex in appearance, equation (16.2) allows one to readily compute \( E y \). For large values of \( (x'_i \beta - d_i)/\sigma \), we see that \( \lambda_i \) is close to 0 and \( \Phi \left( (x'_i \beta - d_i)/\sigma \right) \) is close to 1. We interpret this to mean, for large values of the systematic component \( x'_i \beta \), that the regression function \( E y \) tends to be linear and the usual interpretations apply. The tobit model specification has the greatest impact on observations close to the limiting value \( d_i \).

Equation (16.2) shows that if an analyst ignores the effects of censoring, then the regression function can be quite different than the typical linear regression function, \( E y = x' \beta \), resulting in biased estimates of coefficients. The other tempting path is to exclude limited observations \( y_i = d_i \) from the dataset and again run ordinary regression. However, standard calculations also show that

\[
E (y_i | y_i > d_i) = x'_i \beta + \sigma \frac{\phi \left( (x'_i \beta - d_i)/\sigma \right)}{1 - \Phi \left( (x'_i \beta - d_i)/\sigma \right)},
\]

(16.3)

Thus, this procedure also results in biased regression coefficients.

A commonly used method of estimating the tobit model is maximum likelihood. Employing the normality assumption, standard calculations show that the log-likelihood can be expressed as

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

(16.4)
where \( \{ i : y_i = d_i \} \) and \( \{ i : y_i > d_i \} \) means the sum over the censored and non-censored observations, respectively. Many statistical software packages can readily compute the maximum likelihood estimators, \( b_{MLE} \) and \( s_{MLE} \), as well as corresponding standard errors. Section 11.9 introduces likelihood inference.

For some users, it is convenient to have an algorithm that does not rely on specialized software. A two-stage algorithm due to Heckman (1976) fulfills this need. For this algorithm, first subtract \( d_i \) from each \( y_i \), so that one may take \( d_i \) to be zero without loss of generality. Even for those who wish to use the more efficient maximum likelihood estimators, Heckman’s algorithm can be useful in the model exploration stage as one uses linear regression to help select the appropriate form of the regression equation.

**Heckman’s Algorithm for Estimating Tobit Model Parameters**

(i) For the first stage, define the binary variable

\[
    r_i = \begin{cases} 
    1 & \text{if } y_i > 0 \\
    0 & \text{if } y_i = 0 
\end{cases},
\]

indicating whether or not the observation is censored. Run a probit regression using \( r_i \) as the dependent variable and \( x_i \) as explanatory variables. Call the resulting regression coefficients \( g_{PROBIT} \).

(ii) For each uncensored observation, compute the estimated variable

\[
    \hat{\lambda}_i = \frac{\phi(x_i' g_{PROBIT})}{\Phi(x_i' g_{PROBIT})},
\]

an inverse Mill’s ratio. With this, run a regression of \( y_i \) on \( x_i \) and \( \hat{\lambda}_i \). Call the resulting regression coefficients \( b_{2SLS} \).

The idea behind this algorithm is that equation (16.1) has the same form as the probit model; thus, consistent estimates of the regression coefficients (up to scale) can be computed. The regression coefficients \( b_{2SLS} \) provide consistent and asymptotically normal estimates of \( \beta \). They are, however, inefficient compared to the maximum likelihood estimators, \( b_{MLE} \). Standard calculations (see Exercise 16.1) show that \( \text{Var}(y_i | y_i > d_i) \) depends on \( i \) (even when \( d_i \) is constant). Thus, it is customary to use heteroscedasticity-consistent standard errors for \( b_{2SLS} \).

### 16.3 Application: Medical Expenditures

This section considers data from the Medical Expenditure Panel Survey (MEPS) that were introduced in Section 11.4. Recall that MEPS is a probability survey that provides nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian population. We consider MEPS data from the first panel of 2003 and take a random sample of \( n = 2,000 \) individuals between ages 18 and 65. Section 11.4 analyzed the frequency component, trying to understand the determinants that influenced whether or not people were hospitalized. Section 13.4 analyzed the severity component; given that a person was hospitalized, what are the determinants of medical expenditures? This chapter seeks to unify these two components into a single model of healthcare utilization.
Table 16.1 reviews these explanatory variables and provides summary statistics that suggest their effects on expenditures of inpatient visits. The second column, “Average Expend,” displays the average logarithmic expenditure by explanatory variable, treating no expenditures as a zero (logarithmic) expenditure. This would be the primary variable of interest if one did not decompose the total expenditure into a discrete zero and continuous amount.

Examining this overall average (logarithmic) expenditure, we see that females had higher expenditures than males. In terms of ethnicity, native Americans and Asians had the lowest average expenditures. However, these two ethnic groups accounted for only 5.4% of the total sample size. Regarding regions, it appears that individuals from the West had the lowest average expenditures. In terms of education, more educated persons had lower expenditures. This observation supports the theory that more educated persons take more active roles in keeping their health. When it comes to self-rated health status, poorer physical, mental health and activity related limitations led to greater expenditures. Lower income individuals had greater expenditures and those with insurance coverage had greater average expenditures.

Table 16.1 also describes the effects of explanatory variables on the frequency of utilization and average expenditures for those that used inpatient services. As in Table 11.4, the column “Percent Positive Expend” gives the percentage of individuals that had some positive expenditure, by explanatory variable. The column “Average of Pos Expend” gives the average (logarithmic) expenditure in case where there was an expenditure, ignoring the zeros. This is comparable to the median expenditure in Table 13.5 (given in dollars, not log dollars).

To illustrate, consider that females had a higher average expenditures than males by looking at the “Average Expend” column. Breaking this down into frequency and amount of utilization, we see that females had a higher frequency of utilization but, when they had a positive utilization, the average (logarithmic) expenditure was lower than males. An examination of Table 16.1 shows this observation holds true for other explanatory variables. A variable’s effect on overall expenditures may be positive, negative or non-significant; this effect can be quite different when we decompose expenditures into frequency and amount components.

Table 16.2 compares the ordinary least squares (OLS) regression to maximum likelihood estimates for the tobit model. From this table, we can see that there is a substantial agreement among the t-ratios for these fitted models. This agreement comes from examining the sign (positive or negative) and the magnitude (such as exceeding two for statistical significance) of each variable’s t-ratio. The regression coefficients also largely agree in sign. However, it is not surprising that the magnitudes of the regression coefficients differ substantially. This is because, from equation (16.2), we can see that the tobit coefficients measure the marginal change of the expected latent variable $y^*$, not the marginal change of the expected observed variable $y$, as does OLS.
### Table 16.1. Percent of Positive Expenditures and Average Logarithmic Expenditure, by Explanatory Variable

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable Description</th>
<th>Percent of data</th>
<th>Average Percent Positive of Positive Expenditure</th>
<th>Average Logarithmic Expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demography</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>Age in years between 18 to 65 (mean: 39.0)</td>
<td>52.7</td>
<td>0.91</td>
<td>10.7</td>
</tr>
<tr>
<td>GENDER</td>
<td>1 if female</td>
<td>47.3</td>
<td>0.40</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>1 if male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>ASIAN 1 if Asian</td>
<td>4.3</td>
<td>0.37</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>BLACK 1 if Black</td>
<td>14.8</td>
<td>0.90</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>NATIVE 1 if Native</td>
<td>1.1</td>
<td>1.06</td>
<td>13.6</td>
</tr>
<tr>
<td></td>
<td>WHITE Reference level</td>
<td>79.9</td>
<td>0.64</td>
<td>7.5</td>
</tr>
<tr>
<td>Region</td>
<td>NORTHEAST 1 if Northeast</td>
<td>14.3</td>
<td>0.83</td>
<td>10.1</td>
</tr>
<tr>
<td></td>
<td>MIDWEST 1 if Midwest</td>
<td>19.7</td>
<td>0.76</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td>SOUTH 1 if South</td>
<td>38.2</td>
<td>0.72</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>WEST Reference level</td>
<td>27.9</td>
<td>0.46</td>
<td>5.4</td>
</tr>
<tr>
<td>Education</td>
<td>COLLEGE 1 if college or higher degree</td>
<td>27.2</td>
<td>0.58</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>HIGHSCHOOL 1 if high school degree</td>
<td>43.3</td>
<td>0.67</td>
<td>7.9</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>
</tr>
<tr>
<td></td>
<td>ANYLIMIT 1 if any functional or activity limitation</td>
<td>22.3</td>
<td>1.29</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>Reference level is excellent health</td>
<td>77.7</td>
<td>0.50</td>
<td>5.9</td>
</tr>
<tr>
<td>Income</td>
<td>HINCOME 1 if high income</td>
<td>31.6</td>
<td>0.47</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>MINCOME 1 if middle income</td>
<td>29.9</td>
<td>0.61</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>LINCOME 1 if low income</td>
<td>15.8</td>
<td>0.73</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>NPOOR 1 if near poor</td>
<td>5.8</td>
<td>0.78</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>Reference level is poor/negative</td>
<td>17.0</td>
<td>1.06</td>
<td>13.0</td>
</tr>
<tr>
<td>Insurance</td>
<td>INSURE 1 if covered by public or private health insurance</td>
<td>77.8</td>
<td>0.80</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>0 if have not health insurance in 2003</td>
<td>22.3</td>
<td>0.23</td>
<td>3.1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100.0</td>
<td>0.67</td>
<td>7.9</td>
</tr>
</tbody>
</table>

**Note:** Various levels of health status and activity limitations are included, along with categorical variables for age, gender, ethnicity, region, education, and income level. The table also includes insurance coverage status, with percentages and average logarithmic expenditures for each category.
Table 16.2. Comparison of OLS, Tobit MLE and Two-Stage Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Parameter Estimate</th>
<th>OLS t-ratio</th>
<th>Tobit MLE t-ratio</th>
<th>Two-Stage t-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.123</td>
<td>-0.525</td>
<td>-33.016</td>
<td>2.760</td>
</tr>
<tr>
<td>AGE</td>
<td>0.001</td>
<td>0.091</td>
<td>-0.006</td>
<td>-0.118</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.379</td>
<td>3.711</td>
<td>5.727</td>
<td>4.107</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.115</td>
<td>-0.459</td>
<td>-1.732</td>
<td>-0.480</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.054</td>
<td>0.365</td>
<td>0.025</td>
<td>0.015</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.350</td>
<td>0.726</td>
<td>3.745</td>
<td>0.723</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>1.702</td>
<td>3.828</td>
<td>1.849</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.255</td>
<td>1.693</td>
<td>3.459</td>
<td>1.790</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.146</td>
<td>1.133</td>
<td>1.805</td>
<td>1.056</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>-0.014</td>
<td>-0.089</td>
<td>0.628</td>
<td>0.329</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>-0.027</td>
<td>-0.209</td>
<td>-0.030</td>
<td>-0.019</td>
</tr>
<tr>
<td>POOR</td>
<td>2.297</td>
<td>7.313</td>
<td>13.352</td>
<td>4.436</td>
</tr>
<tr>
<td>FAIR</td>
<td>-0.001</td>
<td>-0.004</td>
<td>1.354</td>
<td>0.528</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.188</td>
<td>1.346</td>
<td>2.740</td>
<td>1.480</td>
</tr>
<tr>
<td>VGOD</td>
<td>0.084</td>
<td>0.622</td>
<td>1.506</td>
<td>0.815</td>
</tr>
<tr>
<td>MNHIPOOR</td>
<td>0.000</td>
<td>-0.001</td>
<td>-0.482</td>
<td>-0.211</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.415</td>
<td>3.103</td>
<td>4.695</td>
<td>3.900</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.482</td>
<td>-2.716</td>
<td>-6.575</td>
<td>-3.035</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.309</td>
<td>-1.868</td>
<td>-4.359</td>
<td>-2.241</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.175</td>
<td>-0.976</td>
<td>-3.414</td>
<td>-1.619</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.116</td>
<td>-0.478</td>
<td>-2.274</td>
<td>-0.790</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.594</td>
<td>4.486</td>
<td>8.534</td>
<td>4.130</td>
</tr>
</tbody>
</table>

Inverse Mill’s Ratio $\hat{\lambda}$ = -3.616 - 0.642
Scale $\sigma^2$ = 4.999 - 14.738

* Two-stage t-ratios are calculated using heteroscedasticity-consistent standard errors.

Table 16.2 also reports the fit using the two-stage Heckman algorithm. The coefficient associated with the inverse Mill’s ratio selection correction is statistically insignificant. Thus, there is general agreement between the OLS coefficients and those estimated using the two-stage algorithm. The two-stage t-ratios were calculated using heteroscedasticity-consistent standard errors, described in Section 5.7.2. Here, we see some disagreement between the t-ratios calculated using Heckman’s algorithm and the maximum likelihood values calculated using the tobit model. For example, GENDER, POOR, HINCOME and MINCOME are statistically significant in the tobit model but are not in the two-stage algorithm. This is troubling because both techniques yield consistent estimators providing the assumptions of the tobit model are valid. Thus, we suspect the validity of the model assumptions for these data; the next section provides an alternative model that turns out to be more suitable for this dataset.

### 16.4 Two-Part Model

One drawback of the tobit model is its reliance on the normality assumption of the latent response. A second, and more important, drawback is that a single latent variable dictates both the magnitude of the response as well as the censoring. As pointed out by Cragg (1971), there are many instances where the limiting amount represents a choice or activity that is separate from the magnitude. For example,
Frequency-Severity Models

in a population of smokers, zero cigarettes consumed during a week may simply represent a lower bound (or limit) and may be influenced by available time and money. However, in a general population, zero cigarettes consumed during a week can indicate that a person is a non-smoker, a choice that could be influenced by other lifestyle decisions (where time and money may or may not be relevant). As another example, when studying healthcare expenditures, a zero represents a person’s choice or decision not to utilize healthcare during a period. For many studies, the amount of healthcare expenditure is strongly influenced by a healthcare provider (such as a physician); the decision to utilize and the amount of healthcare can involve very different considerations.

In the traditional actuarial literature (see for example Bowers et al. 1997, Chapter 2), the individual risk model decomposes a response, typically an insurance claim, into frequency (number) and severity (amount) components. Specifically, let \( r_i \) be a binary variable indicating whether or not the \( i \)th subject has an insurance claim and \( y_i \) describe the amount of the claim. Then, the claim is modeled as

\[
(\text{claim recorded})_i = r_i \times y_i.
\]

This is the basis for the two-part model, where we also use explanatory variables to understand the influence of each component.

\begin{center}
\begin{definition}
\textbf{Two-Part Model}
\begin{enumerate}
\item 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.
\item 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 and gamma regression models.
\end{enumerate}
\end{definition}
\end{center}

Unlike the tobit, in the two-part model one need not have the same set of explanatory variables influencing the frequency and amount of response. However, there is usually overlap in the sets of explanatory variables, where variables are members of both \( x_1 \) and \( x_2 \). Typically, one assumes that \( \beta_1 \) and \( \beta_2 \) are not related so that the joint likelihood of the data can be separated into two components and run separately, as described above.

\begin{center}
\textbf{Example: MEPS Expenditure Data - Continued.} Consider the Section 16.3 MEPS expenditure data using a probit model for the frequency and a linear regression model for the severity. Table 16.3 shows the results from using all explanatory variables to understand their influence on (i) the decision to seek healthcare (frequency) and (ii) the amount of healthcare utilized (severity). Unlike the Table 16.2 tobit model, the two-part models allows each variable to have a separate influence on frequency and severity. To illustrate, the full model results in Table 16.3 show that COLLEGE has no significant impact on frequency but a strong positive impact on severity.

Because of the flexibility of the two-part model, one can also reduce the model
complexity for each component by removing extraneous variables. Table 16.3 shows a reduced model, where age and mental health status variables have been removed from the frequency component; regional, educational, physical status and income variables have been removed from the severity component.

Table 16.3. Comparison of Full and Reduced Two-Part Models

<table>
<thead>
<tr>
<th>Effect</th>
<th>Full Model</th>
<th>Reduced Model</th>
<th>Full Model</th>
<th>Reduced Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Severity</td>
<td>Frequency</td>
<td>Severity</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td>t-ratio</td>
<td>Estimate</td>
<td>t-ratio</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.001</td>
<td>-0.154</td>
<td>0.012</td>
<td>1.368</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.395</td>
<td>4.176</td>
<td>-0.104</td>
<td>-0.469</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.108</td>
<td>-0.429</td>
<td>-0.397</td>
<td>-0.641</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.008</td>
<td>0.062</td>
<td>0.088</td>
<td>0.362</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.284</td>
<td>0.778</td>
<td>-0.639</td>
<td>-0.905</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.283</td>
<td>1.958</td>
<td>-0.649</td>
<td>-2.035</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.239</td>
<td>1.765</td>
<td>0.016</td>
<td>0.052</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.132</td>
<td>1.099</td>
<td>-0.078</td>
<td>-0.294</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.048</td>
<td>0.356</td>
<td>-0.597</td>
<td>-2.066</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>0.002</td>
<td>0.017</td>
<td>-0.415</td>
<td>-1.745</td>
</tr>
<tr>
<td>POOR</td>
<td>0.955</td>
<td>4.576</td>
<td>0.597</td>
<td>1.594</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.087</td>
<td>0.486</td>
<td>-0.211</td>
<td>-0.527</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.184</td>
<td>1.422</td>
<td>0.145</td>
<td>0.502</td>
</tr>
<tr>
<td>VGGOOD</td>
<td>0.095</td>
<td>0.736</td>
<td>0.373</td>
<td>1.233</td>
</tr>
<tr>
<td>MNHPoop</td>
<td>-0.027</td>
<td>-0.164</td>
<td>-0.176</td>
<td>-0.579</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.318</td>
<td>2.941</td>
<td>0.235</td>
<td>0.981</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.468</td>
<td>-3.131</td>
<td>0.490</td>
<td>1.531</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.314</td>
<td>-2.318</td>
<td>0.472</td>
<td>1.654</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.241</td>
<td>-1.626</td>
<td>0.550</td>
<td>1.812</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.145</td>
<td>-0.716</td>
<td>0.067</td>
<td>0.161</td>
</tr>
<tr>
<td>INSURE</td>
<td>0.580</td>
<td>4.154</td>
<td>1.293</td>
<td>3.944</td>
</tr>
</tbody>
</table>

Scale $\sigma^2$ 1.249 1.333

Tobit Type II Model

To connect the tobit and two-part models, let us assume that the frequency is represented by a probit model and use

$$r_i^* = \mathbf{x}'_i \beta_1 + \eta_i$$

to be the latent tendency to be observed. Define $r_i = 1(r_i^* > 0)$ to be the binary variable indicating that an amount has been observed. For the severity component, define

$$y_i^* = \mathbf{x}'_i \beta_2 + \eta_2i$$

to be the latent amount variable. The “observed” amount is

$$y_i = \begin{cases} 
  y_i^* & \text{if } r_i = 1 \\
  0 & \text{if } r_i = 0 
\end{cases}$$

Because responses are censored, the analyst is aware of the subject $i$ and has covariate information even when $r_i = 0$.

If $\mathbf{x}_i = \mathbf{x}_{2i}$, $\beta_1 = \beta_2$ and $\eta_{1i} = \eta_{2i}$, then this is the tobit framework with $d_i = 0$. If $\beta_1$ and $\beta_2$ are not related and if $\eta_{1i}$ and $\eta_{2i}$ are independent, then this is the
two-part framework. For the two-part framework, the likelihood of the observed responses \( \{r_i, y_i\} \) is given by

\[
L = \prod_{i=1}^{n} \left\{ (p_i)^{r_i} (1 - p_i)^{1-r_i} \right\} \prod_{r_i=1} \phi \left( \frac{y_i - \mathbf{x}_i \beta_2}{\sigma_2} \right),
\]

(16.5)

where \( p_i = \Pr (r_i = 1) = \Pr (\mathbf{x}_i \beta_1 + \eta_i > 0) = 1 - \Phi (-\mathbf{x}_i \beta_1) = \Phi (\mathbf{x}_i \beta_1) \). Assuming that \( \beta_1 \) and \( \beta_2 \) are not related, one can separately maximize these two pieces of the likelihood function.

In some instances, it is sensible to assume that the frequency and severity components are related. The tobit model considers a perfect relationship (with \( \eta_i = \eta_{2i} \)) whereas the two-part models assumes independence. For an intermediate model, the tobit type II model allows for a non-zero correlation between \( \eta_{1i} \) and \( \eta_{2i} \). See Amemiya (1985) for additional details. Hsiao et al. (1990) provide an application of the tobit type II model to Canadian collision coverage of private passenger automobile experience.

16.5 Aggregate Loss Model

We now consider two-part models where the frequency may exceed one. For example, if we are tracking automobile accidents, a policyholder may have more than one accident within a year. As another example, we may be interested in the claims for a city or a state and expect many claims per government unit.

To establish notation, for each \( \{i\} \), the observable responses consist of:

- \( N_i \) – the number of claims (events), and
- \( y_{ij}, \ j = 1, ..., N_i \) – the amount of each claim (loss).

By convention, the set \( \{y_{ij}\} \) is empty when \( N_i = 0 \). If one uses \( N_i \) as a binary variable, then this framework reduces to the two-part set-up.

Although we have detailed information on losses per event, the interest often is in aggregate losses, \( S_i = y_{i1} + ... + y_{iN_i} \). In traditional actuarial modeling, one assumes that the distribution of losses are, conditional on the frequency \( N_i \), identical and independent over replicates \( j \). This representation is known as the collective risk model, see, for example, Klugman et al. (2008). We also maintain this assumption.

Data are typically available in two forms:

(i) \( \{N_i, y_{i1}, ..., y_{iN_i}\} \), so that detailed information about each claim is available. For example, when examining personal automobile claims, losses for each claim are available. Let \( \mathbf{y}_i = (y_{i1}, ..., y_{iN_i})' \) be the vector of individual losses.

(ii) \( \{N_i, S_i\} \), so that only aggregate losses are available. For example, when examining losses at the city level, only aggregate losses are available.

We are interested in both forms. Because there are multiple responses (events) per subject \( \{i\} \), one might approach the analysis using multilevel models as described in, for example, Raudenbush and Bryk (2002). Unlike a multilevel structure, we consider data where the number of events are random that we wish to model stochastically and thus use an alternative framework. When only \( \{S_i\} \) is available, the Tweedie GLM introduced in Section 13.6 may be used.
To see how to model these data, consider the first data form. Suppressing the \( \{i\} \) subscript, we decompose the joint distribution of the dependent variables as:

\[
f (N, y) = f (N) \times f (y|N) = \text{frequency } \times \text{conditional severity},
\]

where \( f (N, y) \) denotes the joint distribution of \((N, y)\). This joint distribution equals the product of the two components:

(i) claims frequency: \( f (N) \) denotes the probability of having \( N \) claims; and

(ii) conditional severity: \( f (y|N) \) denotes the conditional density of the claim vector \( y \) given \( N \).

We represent the frequency and severity components of the aggregate loss model as follows.

**Definition. Aggregate Loss Model I**

(i) Use a count regression model with \( N_i \) 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 Poisson and negative binomial models.

(ii) Conditional on \( N_i > 0 \), use a regression model with \( y_{ij} \) as the dependent variable and \( x_{2j} \) as the set of explanatory variables. Denote the corresponding set of regression coefficients as \( \beta_2 \). Typical models include the linear regression, gamma regression and mixed linear models. For the mixed linear models, one uses a subject-specific intercept to account for the heterogeneity among subjects.

To model the second data form, the set-up is similar. The count data model in step 1 will not change. However, the regression model in step 2 will use \( S_i \) as the dependent variable. Because the dependent variable is the sum over \( N_i \) independent replicates, it may be that you will need to allow the variability to depend on \( N_i \).

**Example: MEPS Expenditure Data - Continued.** To get a sense of the empirical observations for claim frequency, we present the overall claim frequency. According to this table, there were a total of 2,000 observations of which 92.15% did not have any claims. There are a total of 203 (= 1 \times 130 + 2 \times 19 + 3 \times 2 + 4 \times 3 + 5 \times 2 + 6 \times 0 + 7 \times 1) claims.

<table>
<thead>
<tr>
<th>Frequency of Claims</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count 0 1 2 3 4 5 6 7 Total</td>
</tr>
<tr>
<td>Number 1,843 130 19 2 3 2 0 1 2,000</td>
</tr>
<tr>
<td>Percentage 92.15 6.50 0.95 0.10 0.15 0.10 0.00 0.10 100.00</td>
</tr>
</tbody>
</table>

Table 16.4 summarizes the regression coefficient parameter fits using the negative binomial model. The results are comparable to the fitted probit models in Table 16.3, where many of the covariates are statistically significant predictors of claim frequency.
This fitted frequency model is based on \( n = 2,000 \) persons. The Table 16.4 fitted severity models are based on \( n_1 + \ldots + n_{2000} = 203 \) claims. The gamma regression model is based on a logarithmic link

\[
\mu_i = \exp (x_i' \beta_2).
\]

Table 16.4 shows that the results from fitting an ordinary regression model are similar to those from fitting the gamma regression model. They are similar in the sense that the sign and statistical significance of coefficients for each variable are comparable. As discussed in Chapter 13, the advantage of the ordinary regression model is its relatively simplicity involving ease of implementation and interpretation. In contrast, the gamma regression model can be a better model for fitting long-tail distributions such as medical expenditures.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Negative Binomial Frequency Parameter Estimate</th>
<th>Ordinary Regression Severity Parameter Estimate</th>
<th>Gamma Regression Severity Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( t )-ratio</td>
<td>( t )-ratio</td>
<td>( t )-ratio</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.005 -0.756</td>
<td>-0.006 -0.747</td>
<td>-0.011 -1.971</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.617 3.351</td>
<td>-0.385 -1.952</td>
<td>-0.826 -4.780</td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.153 -0.306</td>
<td>-0.340 -0.588</td>
<td>-0.711 -1.396</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.144 0.639</td>
<td>0.146 0.686</td>
<td>-0.058 -0.297</td>
</tr>
<tr>
<td>NATIVE</td>
<td>0.445 0.634</td>
<td>-0.331 -0.465</td>
<td>-0.512 -0.841</td>
</tr>
<tr>
<td>NORTHEAST</td>
<td>0.492 1.683</td>
<td>-0.547 -1.792</td>
<td>-0.418 -1.602</td>
</tr>
<tr>
<td>MIDWEST</td>
<td>0.619 2.314</td>
<td>0.303 1.070</td>
<td>0.589 2.234</td>
</tr>
<tr>
<td>SOUTH</td>
<td>0.391 1.603</td>
<td>0.108 0.424</td>
<td>0.302 1.318</td>
</tr>
<tr>
<td>COLLEGE</td>
<td>0.023 0.089</td>
<td>-0.789 -2.964</td>
<td>-0.826 -3.335</td>
</tr>
<tr>
<td>HIGHSCHOOL</td>
<td>-0.085 -0.399</td>
<td>-0.722 -3.396</td>
<td>-0.742 -4.112</td>
</tr>
<tr>
<td>POOR</td>
<td>1.927 5.211</td>
<td>0.664 1.964</td>
<td>0.299 0.989</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.226 0.627</td>
<td>-0.188 -0.486</td>
<td>0.080 0.240</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.385 1.483</td>
<td>0.223 0.802</td>
<td>0.185 0.735</td>
</tr>
<tr>
<td>VGOOD</td>
<td>0.348 1.349</td>
<td>0.429 1.511</td>
<td>0.184 0.792</td>
</tr>
<tr>
<td>MNHPoor</td>
<td>-0.177 -0.583</td>
<td>-0.221 -0.816</td>
<td>-0.470 -1.877</td>
</tr>
<tr>
<td>ANYLIMIT</td>
<td>0.714 3.499</td>
<td>0.579 2.720</td>
<td>0.792 4.171</td>
</tr>
<tr>
<td>HINCOME</td>
<td>-0.622 -2.139</td>
<td>0.723 2.517</td>
<td>0.557 2.290</td>
</tr>
<tr>
<td>MINCOME</td>
<td>-0.482 -1.831</td>
<td>0.720 2.768</td>
<td>0.694 3.148</td>
</tr>
<tr>
<td>LINCOME</td>
<td>-0.460 -1.611</td>
<td>0.631 2.241</td>
<td>0.889 3.693</td>
</tr>
<tr>
<td>NPOOR</td>
<td>-0.465 -1.131</td>
<td>-0.056 -0.135</td>
<td>0.217 0.619</td>
</tr>
<tr>
<td>INSURE</td>
<td>1.312 4.207</td>
<td>1.500 4.551</td>
<td>1.380 4.912</td>
</tr>
<tr>
<td>Dispersion</td>
<td>2.177 1.314</td>
<td>1.131</td>
<td></td>
</tr>
</tbody>
</table>

16.6 Further Reading and References

Property and Casualty

There is a rich literature on modeling the joint frequency and severity distribution of automobile insurance claims. To distinguish this modeling from classical risk theory applications (see, for example, Klugman et al., 2008), we focus on cases where explanatory variables, such as policyholder characteristics, are available. There has
been substantial interest in statistical modeling of claims frequency yet the literature
on modeling claims severity, especially in conjunction with claims frequency, is less
extensive. One possible explanation, noted by Coutts (1984), is that most of the
variation in overall claims experience may be attributed to claim frequency (at least
when inflation was small). Coutts (1984) also remarks that the first paper to analyze
claim frequency and severity separately seems to be Kahane and Levy (1975).

Brockman and Wright (1992) provide an early overview of how statistical modeling
of claims and severity can be helpful for pricing automobile coverage. For compu-
tational convenience, they focused on categorical pricing variables to form cells that
could be used with traditional insurance underwriting forms. Renshaw (1994) shows
how generalized linear models can be used to analyze both the frequency and sever-
ity portions based on individual policyholder level data. Hsiao et al. (1990) note
the “excess” number of zeros in policyholder claims data (due to no claims) and
compare and contrast Tobit, two-part and simultaneous equation models, building
on the work of Weisberg and Tomberlin (1982) and Weisberg et al. (1984). All of
these papers use grouped data, not individual level data in this chapter.

At the individual policyholder level, Frangos and Vrontos (2001) examined a
claim frequency and severity model, using negative binomial and Pareto distribu-
tions, respectively. They used their statistical model to develop experience rated
approach, fitting not only cross-sectional data but also following policyholders over
time. Pinquet was interested in two lines of business, claims at fault and not at fault
with respect to a third party. For each line, Pinquet hypothesized a frequency and
severity component that were allowed to be correlated to one another. In particular,
the claims frequency distribution was assumed to be bivariate Poisson. Severities
were modeled using lognormal and gamma distributions.

Healthcare

The two-part model became prominent in the healthcare literature upon adoption
by Rand Health Insurance Experiment researchers (Duan et al, 1983, Manning et
al, 1987). They used the two-part model to analyze health insurance cost sharing’s
effect on healthcare utilization and expenditures because of the close resemblance
of the demand for medical care to the two decision-making processes. That is, the
amount of healthcare expenditures is largely unaffected by an individual’s decision
to seek treatment. This is because physicians, as the patients’ (principal) agents,
would tend to decide the intensity of treatments as suggested by the principal-agent

The two-part model has become widely used in the healthcare literature despite
some criticisms. For example, Maddala (1985) argued that two-part modeling is
not appropriate for non-experimental data because individuals’ self-selection into
different health insurance plans is an issue. (In the Rand Health Insurance Ex-
periment, the self-selection aspect was not an issue because participants were ran-
domly assigned to health insurance plans.) See Jones (2000) and Mullahy (1998) for
overviews.

Two-part models remain attractive in modeling healthcare usage because they
provide insights into the determinants of initiation and level of healthcare usage.
The decision to utilize healthcare by individuals is related primarily to personal
characteristics whereas the cost per user may be more related to characteristics of the healthcare provider.

Chapter References


Assume that \( y \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \). Let \( \phi(.) \) and \( \Phi(.) \) be the standard normal density and distribution functions, respectively. Define \( h(d) = \phi(d) / (1 - \Phi(d)) \), a hazard rate. Let \( d \) be a known constant and \( d_s = (d - \mu) / \sigma \) be the standardized version.

a. Determine the density of \( y \), conditional on \( \{ y > d \} \)

b. Show that \( E (y|y > d) = \mu + \sigma h(d_s) \).

c. Show that \( E (y|y \leq d) = \mu - \sigma \phi(d) / \Phi(d) \).

d. Show that \( \text{Var} (y|y > d) = \sigma (1 - \delta(d_s)) \), where \( \delta(d) = h(d) (h(d) - d) \).

e. Show that \( E \ max (y, d) = \mu + d - ((\mu + \sigma h(d_s)) (1 - \Phi(d_s)) + d \Phi(d_s)) \).

f. Show that \( E \ min (y, d) = \mu + d - ((\mu + \sigma h(d_s)) (1 - \Phi(d_s)) + d \Phi(d_s)) \).

Verify the log-likelihood in equation (16.4) for the tobit model.

Verify the log-likelihood in equation (16.5) for the two-part model.

Derive the log-likelihood for the tobit type two model. Show that your log-likelihood reduces to equation (16.5) in the case of uncorrelated disturbance terms.