Instructors' Manual for
Regression Modeling with Actuarial and Financial Applications

Edward W. Frees

Data Sets

Table 1. Anscombe's Data

The data is due to Anscombe (1973). The purpose of dealing with this data set is to demonstrate how plotting data can reveal important information that is not evident in numerical summary statistics.

File Name: AnscombeData	Number of obs: 11	Number of variables: 7
Variable	Number of Obs Missing	**Description**
ObsNum		The number of the observation
x1		Generic explanatory variable
y1		Generic dependent variable
y2		Version 2 of the dependent variable
y3		Version 3 of the dependent variable
x2		Version 2 of the explanatory variable
y4		Version 4 of the dependent variable

Source: Anscombe (1973).

Table 1

```
Example of the first five observations:
  ObsNum x1   y1   y2    y3 x2   y4
1      1 10 8.04 9.14  7.46  8 6.58
2      2  8 6.95 8.14  6.77  8 5.76
3      3 13 7.58 8.74 12.74  8 7.71
4      4  9 8.81 8.77  7.11  8 8.84
5      5 11 8.33 9.26  7.81  8 8.47
```

Table 2. Automobile Bodily Injury Claims

We consider automobile injury claims data using data from the Insurance Research Council (IRC), a division of the American Institute for Chartered Property Casualty Underwriters and the Insurance Institute of America. The data, collected in 2002, contains information on demographic information about the claimant, attorney involvement and the economic loss (LOSS, in thousands), among other variables. We consider here a sample of $n = 1,340$ losses from a single state. The full 2002 study contains over 70,000 closed claims based on data from thirty-two insurers. The IRC conducted similar studies in 1977, 1987, 1992 and 1997.

File Name: AutoBI	Number of obs: 1340	Number of variables: 8
Variable	Number of Obs Missing	**Description**
CASENUM		Case number to identify the claim
ATTORNEY		Whether the claimant is represented by an attorney (=1 if yes and =2 if no)
CLMSEX	12	Claimant's gender (=1 if male and =2 if female)
MARITAL	16	claimant's marital status (=1 if married, =2 if single, =3 if widowed, and =4 if divorced/separated)
CLMINSUR	41	Whether or not the driver of the claimant's vehicle was uninsured (=1 if yes, =2 if no, and =3 if not applicable)
SEATBELT	48	Whether or not the claimant was wearing a seatbelt/child restraint (=1 if yes, =2 if no, and =3 if not applicable)
CLMAGE	189	Claimant's age
LOSS		The claimant's total economic loss (in thousands)

Source: Insurance Research Council (IRC).

Table 2

```
Example of the first five observations:
  CASENUM ATTORNEY CLMSEX MARITAL CLMINSUR SEATBELT CLMAGE   LOSS
1       5        1      1      NA        2        1     50 34.940
2      13        2      2       2        1        1     28 10.892
3      66        2      1       2        2        1      5  0.330
4      71        1      1       1        2        2     32 11.037
5      96        2      1       4        2        1     30  0.138
```

Table 3. Automobile UK Collision Claims

This data is due to Mildenhall (1999). Mildenhall (1999) considered 8,942 collision losses from private passenger United Kingdom (UK) automobile insurance policies. The data were derived from Nelder and McCullagh (1989, Section 8.4.1) but originated from Baxter et al. (1980). We consider here a sample of $n = 32$ of Mildenhall data for eight driver types (age groups) and four vehicle classes (vehicle use). The average severity is in pounds sterling adjusted for inflation.

File Name: AutoCollision	Number of obs: 32	Number of variables: 4
Variable	Number of Obs Missing	**Description**
Age		Age of driver
Vehicle_Use		Purpose of the vehicle use: "DriveShort" means drive to work but less than 10 miles "DriveLong" means drive to work but more than 10 miles
Severity		Average amount of claims (in pounds sterling)
Claim_Count		Number of claims

Source: Mildenhall (1999).

Table 3

```
Example of the first five observations:
    Age Vehicle_Use Severity Claim_Count
1 17-20 Pleasure     250.48          21
2 17-20 DriveShort   274.78          40
3 17-20 DriveLong    244.52          23
4 17-20 Business     797.8            5
5 21-24 Pleasure     213.71          63
```

Table 4. Automobile Insurance Claims

We examine claims experience from a large midwestern (US) property and casualty insurer for private passenger automobile insurance. The dependent variable is the amount paid on a closed claim, in (US) dollars (claims that were not closed by year end are handled separately). Insurers categorize policyholders according to a risk classification system. This insurer's risk classification system is based on automobile operator characteristics and vehicle characteristics, and these factors are summarized by the risk class categorical variable CLASS.

File Name: Autoclaims	Number of obs: 6773	Number of variables: 5
Variable	Number of Obs Missing	**Description**
STATE		Codes 01 to 17 used, with each code randomly assigned to an actual individual state
CLASS		Rating class of operator, based on age, gender, marital status, use of vehicle, as coded in a separate PDF file
GENDER		Gender of operator
AGE		Age of operator
PAID		Amount paid to settle and close a claim

Source: Insurance company data provided as a personal communication to the author.

Table 4

```
Example of the first five observations:
  State_Code Class Gender Age    Paid
1   STATE 14   C6       M  97 1134.44
2   STATE 15   C6       M  96 3761.24
3   STATE 15   C11      M  95 7842.31
4   STATE 15   F6       F  95 2384.67
5   STATE 15   F6       M  95  650.00
```

Table 5. Capital Asset Pricing Model

We study a financial application, the Capital Asset Pricing Model, often referred to by the acronym CAPM. The name is something of a misnomer in that the model is really about returns based on capital assets, not the prices themselves. The types of assets that we examine are equity securities that are traded on an active market, such as the New York Stock Exchange (NYSE).

An intuitively appealing idea, and one of the basic characteristics of the CAPM, is that there should be a relationship between the performance of a security and the market. One rationale is simply that if economic forces are such that the market improves, then those same forces should act upon an individual stock, suggesting that it also improve. Another rationale for a relationship between security and market returns comes from financial economics theory. Other things equal, investors would like to select a return with a high expected value and low standard deviation, the latter being a measure of riskiness.

Testing economic theory, or models arising from any discipline, involves collecting data. The CAPM theory is about ex-ante (before the fact) returns even though we can only test with ex-post (after the fact) returns. Before the fact, the returns are unknown and there is an entire distribution of returns. After the fact, there is only a single realization of the security and market return. Because at least two observations are required to determine a line, CAPM models are estimated using security and market data gathered over time. In this way, several observations can be made. For the purposes of our discussions, we follow standard practice in the securities industry and examine monthly prices. Specifically, these data consist of monthly returns over the five year period from January, 1986 to December, 1990, inclusive.

File Name: CAPM	Number of obs: 60	Number of variables: 3
Variable	Number of Obs Missing	**Description**
AMERICAN		Monthly returns of American Family Company
LINCOLN		Monthly security returns from the Lincoln National Insurance Corporation
MARKET		Monthly market returns from index of the Standard & Poor's 500 Index

Source: Center for Research on Security Prices, University of Chicago.

Table 5

```
Example of the first five observations:
   AMERICAN   LINCOLN    MARKET
1  0.303167  0.164588  0.004702
2  0.080092  0.030238  0.067982
3 -0.015054 -0.006289  0.052660
4  0.043668 -0.065401 -0.014899
5  0.107950  0.041002  0.045552
```

Table 6. Insurance Redlining

Do insurance companies use race as a determining factor when making insurance available? Fienberg (1985) gathered data from a report issued by the U.S. Commission on Civil Rights about the number of homeowners and residential fire insurance policies issued in Chicago over the months of December 1977 through February 1978. Policies issued were categorized as either part of the standard voluntary market or the substandard, involuntary market. The involuntary market consists of "fair access to insurance requirements" (FAIR) plans; these are state insurance programs sometimes subsidized by private companies. These plans provide insurance to people who would otherwise be denied insurance on their property due to high-risk problems. The main purpose is to understand the relationship between insurance activity and the variable "race", the percentage minority. Data are available for $n = 47$ zip codes in the Chicago area. These data have also been analyzed by Faraway (2005).

To help control for the size of the expected loss, Fienberg also gathered theft and fire data from Chicago's police and fire departments. Another variable that gives some information about loss size is the age of the house. The median income, from the Census Bureau, gives indirect information on the size of the expected loss as well as whether the applicant can afford insurance.

File Name: Chicago	Number of obs: 47	Number of variables: 8
Variable	Number of Obs Missing	**Description**
zipcode		Zip (postal) code
race		Racial composition in percent minority
fire		Fires per 1,000 housing units
theft		Thefts per 1,000 population
age		Percent of housing units built in or before 1939
volact		New homeowner policies plus renewals, minus cancellations and non-renewals per 100 housing units
involact		New FAIR plan policies and renewals per 100 housing units
income		Median family income

Source: Fienberg (1985).

Table 6

```
Example of the first five observations:
  zipcode race fire theft  age volact involact income
1   60626 10.0  6.2    29 60.4    5.3      0.0  11744
2   60640 22.2  9.5    44 76.5    3.1      0.1   9323
3   60613 19.6 10.5    36 73.5    4.8      1.2   9948
4   60657 17.3  7.7    37 66.9    5.7      0.5  10656
5   60614 24.5  8.6    53 81.4    5.9      0.7   9730
```

Table 7. CEO Compensation

The data were drawn from the May 25, 1992 issue of *Forbes Magazine* entitled "What 800 Companies Paid for their Bosses." This article provides several measures of CEO compensation, as well as characteristics of the CEO and measures of his firm's performance. We say "his" because of the 800 CEOs studied in this article, only one was a woman. The data is used to study CEO and firm characteristics to determine the important factors influencing CEO compensation.

To understand the determinants of CEO compensation, one hundred observations were randomly selected from the 800 listed in the *Forbes* article. Although the *Forbes* article did not cite the basis for a firm to be included in its survey, the 800 companies seem to represent the largest publicly traded companies in the United States. Our sample of one hundred CEOs and their firms represent a cross-sectional sample of America's largest corporations. In our cross-section, the CEO and firm characteristics were based on 1991 measures.

File Name: CeoCompensation	Number of obs: 100	Number of variables: 12
Variable	Number of Obs Missing	**Description**
COMP		Sum of salary, bonus and other 1991 compensation, in thousands of dollars. Other compensation does not include stock gains.
AGE		The CEOs age, in years
EDUCATN		The CEOs education level, 1 for no college degree, 2 for a college undergraduate degree and 3 for a graduate degree
BACKGRD		Background type, 0 for unknown, 1 for technical, 2 for insurance, 3 for operations, 4 for banking, 5 for legal, 6 for marketing, 7 for administration, 8 for sales, 9 for financial and 10 for journalism
TENURE		Number of years employed by the firm
EXPER		Number of years as the firm CEO
SALES		1991 sales revenues, in millions of dollars
VAL		Market value of the CEO's stock, in natural logarithmic units
PCNTOWN		Percentage of firm's market value owned by the CEO
PROF		1991 profits of the firm, before taxes, in millions of dollars
COMPANY		Company name
BIRTH		The CEOs birthplace

Source: Forbes Magazine.

Table 7

```
Example of the first five observations:
  COMP AGE EDUCATN BACKGRD TENURE EXPER SALES VAL PCNTOWN PROF COMPANY  BIRTH
1 1948  55       1       1     23  23.0  1227 7.6    0.55  145    AdvM    chi
2  809  59       1       2     38   0.5 19196 0.4    0.01  505   aetna    chi
3  721  53       2       1     26   0.5   839 1.5    0.10  -60   aller   sanf
4 2027  62       2       2     25   5.0  8379 3.4    0.04  806    amer  vertx
5 2094  63       1       3     41   8.0 10818 5.9    0.04 1166   ameri bigrun
```

Table 8. Galton Heights

These data are from Galton's 1885 paper, including the heights of 928 adult children, classified by an index of their parents' height. Here, all female heights were multiplied by 1.08, and the index was created by taking the average of the father's height and rescaled mother's height. Galton was aware that each column could be adequately approximated by a normal curve. In developing regression analysis, he provided a single model for the entire data set.

Galton's 1885 regression data shows that much of the information concerning the height of an adult child can be attributed to, or "explained," in terms of the parents' height.

File Name: Galton	Number of obs: 102	Number of variables: 5
Variable	Number of Obs Missing	**Description**
NUMBER		Number of families in a cell category
MIDPARNT		Category for the height of the midparent that defines the column of the cell
CHILD		Category for the height of the child that defines the row of the cell
PARENTC		Average of father's height and rescaled (multiplied by 1.08) mother's height, that defines the column of the cell
CHILDC		Height of adult child in inches, that defines the row of the cell

Source: Stigler (1986).

Table 8

```
Example of the first five observations:
  NUMBER MIDPARNT CHILD PARENTC CHILDC
1      1        1    12    74.5   72.2
2      3        1    13    74.5   73.2
3      1        2     8    73.5   68.2
4      2        2     9    73.5   69.2
5      1        2    10    73.5   70.2
```

Table 9. MEPS Health Expenditures

The data were 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 of 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. You can learn more about MEPS at http://www.meps.ahrq.gov/mepsweb/.

We consider MEPS data from the panels 7 and 8 of 2003 that consists of 18,735 individuals between ages 18 and 65. From this sample, we took a random sample of 2,000 individuals. From this sample, there are 157 individuals that had positive inpatient expenditures. There are also 1,352 that had positive outpatient expenditures. We will analyze these two samples separately. Our dependent variables consist of amounts of expenditures for inpatient (EXPENDIP) and outpatient (EXPENDOP) visits. For MEPS, 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 in 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.)

File Name: HealthExpend	Number of obs: 2000	Number of variables: 28
Variable	Number of Obs Missing	**Description**
AGE		Age in years between 18 and 65
ANYLIMIT		Any activity limitation (=1 if any functional/activity limitation, =0 if otherwise)
COLLEGE		1 if college or higher degree
HIGHSCH		1 if high school degree
GENDER		Indicate gender of patient (=1 if female, =0 if male)
MNHPOOR		Self-rated mental health (=1 if poor or fair, =0 if good to excellent mental health)
insure		Insurance coverage (=1 if covered by public/private health insurance in any month of 1996, =0 if have no health insurance in 1996)
USC		1 if dissatisfied with one's usual source of care
UNEMPLOY		Employment status of patients
MANAGEDCARE		1 if enrolled in an HMO or gatekeeper plan
famsize		Family size of patients
COUNTIP		Number of inpatient visits
EXPENDIP		Amounts of expenditures for inpatient visits
COUNTOP		Number of outpatient visits
EXPENDOP		Amounts of expenditures for outpatient visits
RACE		Race of patient described by words (Asian, Black, Native, White and other)
RACE1		Race of patient described by numbers (=1 if Asian, =2 if Black, =3 if Native, =4 if White and =0 if others)
REGION		Region of patient described by words (WEST, NORTHEAST, MIDWEST and SOUTH)
REGION1		Region of patient described by numbers (=0 if WEST, =1 if NORTHEAST, =2 if MIDWEST and =3 if SOUTH)
EDUC		Level of education received described by words (LHIGHSC, HIGHSCH and COLLEGE)
EDUC1		Level of education received described by numbers (=0 if lower than high school, =1 if high school and =2 if college)
MARISTAT		Married status of patients described by words (NEVMAR, MARRIED, WIDOWED and DIVSEP)
MARISTAT1		Married status of patients described by words (=0 if never married, =1 if married, =2 if widowed and =3 if divorced or seperated)
INCOME		Income compared to poverty line described by words (POOR, NPOOR, LINCOME, MINCOME and HINCOME)
INCOME1		Income compared to poverty line described by numbers (=0 if poor, =1 if near poor, =2 if low income, =3 if middle income and =4 if high income)
PHSTAT		Self-rated physical health status described by words (EXCE, VGOO, GOOD, FAIR and POOR)
PHSTAT1		Self-rated physical health status described by numbers (=0 if excellent, =1 if very good, =2 if good, =3 if fair and =4 if poor)
INDUSCLASS		Industry each patient belongs to

Source: Medical Expenditure Panel Survey (MEPS).

Table 9

Example of the first five observations:

	AGE	ANYLIMIT	COLLEGE	HIGHSCH	GENDER	MNHPOOR	insure	USC	UNEMPLOY	MANAGEDCARE	famsize
1	30	0	0	0	0	0	0	0	0	0	3
2	56	1	0	1	0	0	1	1	1	1	3
3	55	1	1	0	0	0	1	1	0	0	2
4	47	0	1	0	1	0	1	1	0	0	2
5	50	0	1	0	1	0	1	1	1	1	1

	COUNTIP	EXPENDIP	COUNTOP	EXPENDOP	RACE	RACE1	REGION	REGION1	EDUC	EDUC1
1	0	0.00	0	0.00	WHITE	4	MIDWEST	2	LHIGHSC	0
2	0	0.00	5	2384.56	BLACK	2	SOUTH	3	HIGHSCH	1
3	2	16121.45	42	29729.56	WHITE	4	MIDWEST	2	COLLEGE	2
4	0	0.00	4	110.00	BLACK	2	NORTHEAST	1	COLLEGE	2
5	0	0.00	43	3298.95	WHITE	4	WEST	0	COLLEGE	2

	MARISTAT	MARISTAT1	INCOME	INCOME1	PHSTAT	PHSTAT1	INDUSCLASS
1	MARRIED	1	MINCOME	3	EXCE	0	TRANSINFO
2	MARRIED	1	MINCOME	3	GOOD	2	
3	MARRIED	1	HINCOME	4	EXCE	0	NATRESOURCE
4	MARRIED	1	HINCOME	4	FAIR	3	
5	DIVSEP	3	LINCOME	2	GOOD	2	

Table 10. Hong Kong Horse Racing

The race track is a fascinating example of financial market dynamics at work. From racing forms, newspapers and so on, there are many explanatory variables that are publicly available that might help us predict whether a horse wins. 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 and the odds can be readily converted to the crowd's assessment of the probabilities of winning.

Here 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, the response variable FINISH is the indicator of a horse winning a race and the explanatory variable WIN is the crowd's a priori probability assessment of a horse winning a race.

File Name: HKHorse	Number of obs: 925	Number of variables: 2
Variable	Number of Obs Missing	**Description**
FINISH		Indicator of a horse winning a race
WIN		Crowd's a priori probability assessment of a horse winning a race

Source: Frees (1996).

Table 10

```
Example of the first five observations:
  FINISH        WIN
1      0 0.04387264
2      0 0.10525265
3      0 0.17812790
4      0 0.13731299
5      0 0.02147788
```

Table 11. Hospital Costs

The data were from the Nationwide Inpatient Sample of the Healthcare Cost and Utilization Project (NIS-HCUP), a nationwide survey of hospital costs conducted by the US Agency for Healthcare Research and Quality (AHRQ). We restrict consideration to Wisconsin hospitals and analyze a random sample of $n = 500$ claims from 2003 data. Although the data comes from hospital records, it is organized by individual discharge and so we have information about the age and gender of the patient discharged. Specifically, we consider patients aged 0-17 years. We will use these data to consider the frequency of hospitalization. The data will also be used to model the severity of hospital charges, by age and gender.

File Name: HospitalCosts	Number of obs: 500	Number of variables: 6
Variable	Number of Obs Missing	**Description**
AGE		Age of the patient discharged
FEMALE		Binary variable that indicates if the patient is female
LOS		Length of stay, in days
RACE	1	Race
TOTCHG		Hospital discharge costs
APRDRG		All patients refined diagnosis related group

Source: Nationwide Inpatient Sample of the Healthcare Cost and Utilization Project (NIS-HCUP), conducted by the US Agency for Healthcare Research and Quality (AHRQ).

Table 11

Example of the first five observations:

	AGE	FEMALE	LOS	RACE	TOTCHG	APRDRG
1	17	1	2	1	2660	560
2	17	0	2	1	1689	753
3	17	1	7	1	20060	930
4	17	1	1	1	736	758
5	17	1	1	1	1194	754

Table 12. Initial Public Offering (IPO)

As a financial analyst, one wishes to convince a client of the merits of investing in firms that have just entered a stock exchange, as an IPO (initial public offering). Thus, we gather data on 116 firms that priced during the six-month time frame of January 1, 1998 through June 1, 1998. By looking at this recent historical data, we are able to compute RETURN, the firm's one-year return (in percent).

We are also interested in looking at financial characteristics of the firm that may help us understand (and predict) the return. We initially examine REVENUE, the firm's 1997 revenues in millions of dollars. Unfortunately, this variable was not available for six firms. Thus, only 110 firms that have both REVENUES and RETURNS.

File Name: IPO	Number of obs: 116	Number of variables: 6
Variable	Number of Obs Missing	**Description**
COMPANY		Name of the company
TICKER		Ticker symbol
RETURN		The firm's one-year return (in percent)
REV	6	The firm's revenues in millions of dollars
LnREV	6	Logarithm of revenues
PRICEIPO		Initial price of the stock

Source: .

Table 12

Example of the first five observations:

	COMPANY	TICKER	RETURN	REV	LnREV	PRICEIPO
1	Inktomi Corp.	INKT	4.333333333	5.785000	1.755268	18
2	IBS Interactive	IBSX	2.458333333	2.741000	1.008323	6
3	Frontline Communications Corp.	FCCN	2.09375	0.098699	-2.315680	4
4	MIPS Technologies	MIPS	2.080357143	40.307000	3.696525	14
5	Broadcom Corp.	BRCM	1.921875	36.955000	3.609701	24

Table 13. Stock Market Liquidity

An investor's decision to purchase a stock is generally made with a number of criteria in mind. First, investors usually look for a high expected return. A second criterion is the riskiness of a stock which can be measured through the variability of the returns. Third, many investors are concerned with the length of time that they are committing their capital with the purchase of a security. Many income stocks, such as utilities, regularly return portions of capital investments in the form of dividends. Other stocks, particularly growth stocks, return nothing until the sale of the security. Thus, the average length of investment in a security is another criterion. Fourth, investors are concerned with the ability to sell the stock at any time convenient to the investor. We refer to this fourth criterion as the liquidity of the stock. The more liquid is the stock, the easier it is to sell. To measure the liquidity, in this study we use the number of shares traded on an exchange over a specified period of time (called the VOLUME). We are interested in studying the relationship between the volume and other financial characteristics of a stock.

We begin this study with 123 companies whose options were traded on December 3, 1984. The stock data were obtained from Francis Emory Fitch, Inc. for the period from December 3, 1984 to February 28, 1985.

File Name: Liquidity	Number of obs: 123	Number of variables: 9
Variable	Number of Obs Missing	**Description**
AVGT		Average time between transactions, in minutes
VOLUME		Three months total trading volume, in millions of shares
NTRAN		Three months total number of transactions
PRICE		Opening stock price on January 2, 1985, in U.S. dollars
SHARE		Number of outstanding shares on December 31, 1984, in millions of shares
VALUE		Market equity value obtained by taking the product of PRICE and SHARE
DEBEQ		Debt-to-equity ratio, financial leverage
TIC		Company ticker symbol
COMPANY		Company name

Source: Francis Emory Fitch, Inc., Standard & Poor's Compustat, and University of Chicago's Center for Research on Security Prices.

Table 13

Example of the first five observations:

	AVGT	VOLUME	NTRAN	PRICE	SHARE	VALUE	DEBEQ	TIC	COMPANY
1	3.452	16.221	6273	37.250	81.141	3.002	0.897	AA	ALUMINUM CO AMER
2	13.561	5.693	1548	22.500	27.088	0.609	6.394	AAL	ALEXANDER & ALEX SVCS
3	2.884	11.965	7582	21.000	189.680	4.007	1.792	AEP	AMERICAN ELEC PWR INC
4	5.674	3.834	3771	24.375	13.492	0.325	3.089	AGE	EDWARDS AG INC
5	2.909	13.235	7434	28.750	72.600	2.087	0.644	AHS	AMERICAN HOSP SUPPLY

Table 14. Massachusetts Bodily Injury

Rempala and Derrig (2005) considered claims arising from automobile bodily injury insurance coverages. These are amounts incurred for outpatient medical treatments that arise from automobile accidents, typically sprains, broken collarbones and the like. The data consists of a sample of claims from Massachusetts that were closed in 2001 (by "closed", we mean that the claim is settled and no additional liabilities can arise from the same accident). Rempala and Derrig were interested in developing procedures for handling mixtures of "typical" claims, and those from providers who reported claims fraudulently. For this sample, we consider only those "typical" claims, ignoring the potentially fraudulent ones. Potentially fraudulent claims are from provider=A, our analysis consists of claims from "Other" providers.

File Name: MassBodilyInjury	Number of obs: 348	Number of variables: 5
Variable	Number of Obs Missing	**Description**
Rownum		Identification of the claim
claims		Claims arising from automobile bodily injury insurance coverages
provider		Health care provider is either "A" or "Other"
providerA		Binary variable indicating the presence of "Other" provider
logclaims		Logarithm of claims

Source: Rempala and Derrig (2005).

Table 14

```
Example of the first five observations:
  ID claims provider providerA logclaims
1  1  0.045    Other         1    -3.101
2  2  0.047    Other         1    -3.058
3  3  0.070    Other         1    -2.659
4  4  0.075    Other         1    -2.590
5  5  0.077    Other         1    -2.564
```

Table 15. Insurance Company Expenses

Like every other business, insurance companies seek to minimize expenses associated with doing business in order to enhance profitability. To study expenses, we examine a random sample of 500 insurance companies from the National Association of Insurance Commissioners (NAIC) database of over 3,000 companies. The NAIC maintains one of the world's largest insurance regulatory databases; we consider here data that is based on 2005 annual reports for all the property and casualty insurance companies in United States. The annual reports are financial statements that use statutory accounting principles.

File Name: NAICExpense	Number of obs: 384	Number of variables: 15
Variable	Number of Obs Missing	**Description**
COMPANY_NAME		Name of the company
GROUP		Indicates if the company is affiliated
MUTUAL		Indicates if the company is a mutual company
STOCK		Indicates if the company is a stock company
RBC		Risk-Based Capital
EXPENSES		Total expenses incurred, in millions of dollars
STAFFWAGE		Annual average wage of the insurer's administrative staff, in thousands of dollars
AGENTWAGE	19	Annual average wage of the insurance agent, in thousands of dollars
LONGLOSS		Losses incurred for long tail lines, in millions of dollars
SHORTLOSS		Losses incurred for short tail lines, in millions of dollars
GPWPERSONAL		Gross premium written for personal lines, in millions of dollars
GPWCOMM		Gross premium written for commercial lines, in millions of dollars
ASSETS		Net admitted assets, in millions of dollars
CASH		Cash and invested assets, in millions of dollars
LIQUIDRATIO		The ratio of the liquid assets to the current liabilities level

Source: National Association of Insurance Commissioners (NAIC).

Table 15

Example of the first five observations:

	COMPANY_NAME	GROUP	MUTUAL	STOCK	RBC	EXPENSES
1	Tift Area Captive Ins Co	0	0	1	228184000	0.0008019802
2	Alliance Of Nonprofits For Ins RRG	0	0	0	1627708000	0.0044878635
3	GA Timber Harvesters Mut Captive	0	1	0	422907000	0.0019045075
4	American Natl Lloyds Ins Co	1	0	0	652906000	0.0022909382
5	Chubb Natl Ins Co	1	0	1	8124624000	0.0182956574

	STAFFWAGE	AGENTWAGE	LONGLOSS	SHORTLOSS	GPWPERSONAL	GPWCOMM	ASSETS
1	84.40508	77.46100	0.0001873308	0.000000000	0.00000000	0.001375438	0.002949942
2	81.56754	84.87802	0.0027822909	0.000000000	0.00000000	0.012272512	0.022170349
3	84.40508	77.46100	0.0010121463	0.001329539	0.00000000	0.005028351	0.004617343
4	82.49788	75.71071	0.0000000000	0.002979557	0.02954504	0.001986159	0.043719914
5	79.26495	78.24790	0.0107939577	0.011777314	0.04061412	0.058094479	0.144773034

	CASH	LIQUIDRATIO
1	0.003258406	110.45661
2	0.019760347	89.12961
3	0.003499702	75.79472
4	0.040934885	93.62984
5	0.138424153	95.61460

Table 16. Outlier Example

This data set we considered here is a fictitious data set of 19 points plus three points, labeled A, B, and C. Think of the first 19 points as "good" observations that represent some type of phenomena. We want to investigate the effect of adding a single aberrant point.

File Name: OutlierExample	Number of obs: 22	Number of variables: 3
Variable	Number of Obs Missing	**Description**
X		Explanatory variable
Y		Response variable
CODES		Codes for type of point: 0 if basic, 1 if an outlier, 2 if an influential point that is not an outlier and 3 if an outlier that is influential

Source: Author calculations.

Table 16

```
Example of the first five observations:
    X   Y CODES
1 1.5 3.0     0
2 1.7 2.5     0
3 2.0 3.5     0
4 2.2 3.0     0
5 2.5 3.1     0
```

Table 17. Refrigerator Prices

What characteristics of a refrigerator are important in determining its price (PRICE)? We consider here several characteristics of a refrigerator, including the size of the refrigerator in cubic feet (RSIZE), the size of the freezer compartment in cubic feet (FSIZE), the average amount of money spent per year to operate the refrigerator (ECOST, for "energy cost"), the number of shelves in the refrigerator and freezer doors (SHELVES), and the number of features (FEATURES). The features variable includes shelves for cans, see-through crispers, ice makers, egg racks and so on.

Both consumers and manufacturers are interested in models of refrigerator prices. Other things equal, consumers generally prefer larger refrigerators with lower energy costs that have more features. Due to forces of supply and demand, we would expect consumers to pay more for these refrigerators. A larger refrigerator with lower energy costs that has more features at the similar price is considered a bargain to the consumer. How much extra would the consumer be willing to pay for this additional space? A model of prices for refrigerators on the market provides some insight to this question.

File Name: Refrigerator	Number of obs: 37	Number of variables: 8
Variable	Number of Obs Missing	**Description**
PRICE		Price of a refrigerator
ECOST		Average amount of money spent per year to operate the refrigerator
RSIZE		Size of the refrigerator in cubic feet
FSIZE		Size of the freezer compartment in cubic feet
SHELVES		Number of shelves in refrigerator and freezer doors
S_SQ_FT		Amount of shelf space, measured in square feet
FEATURES		Number of features
BRANDNAM		Brand name of the refrigerator

Source: Consumer Reports, 1992, July. "Refrigerators: A Comprehensive Guide to the Big White Box".

Table 17

```
Example of the first five observations:
  PRICE ECOST RSIZE FSIZE SHELVES S.SQ.FT FEATURES BRANDNAM
1   595    75  12.8   5.7       3    25.4        2  Admiral
2   685    75  12.9   5.7       3    26.7        1  Admiral
3   535    67  13.3   4.5       1    24.0        6    Amana
4   600    67  13.2   4.5       3    23.5        5    Amana
5   605    67  13.3   4.5       3    24.0        3    Amana
```

Table 18. Risk Managers Cost Effectiveness

The data for this study were provided by Professor Joan Schmit and are discussed in more detail in the paper, "Cost effectiveness of risk management practices," Schmit and Roth (1990). The data are from a questionnaire that was sent to 374 risk managers of large U.S.-based organizations. The purpose of the study was to relate cost effectiveness to management's philosophy of controlling the company's exposure to various property and casualty losses, after adjusting for company effects such as size and industry type.

File Name: RiskSurvey	Number of obs: 73	Number of variables: 7
Variable	Number of Obs Missing	**Description**
FIRMCOST		The measure of the firm's risk management cost effectiveness, defined as total property and casualty premiums and uninsured losses as a percentage of total assets
ASSUME		Per occurrence retention amount as a percentage of total assets
CAP		Indicates that the firm owns a captive insurance company
SIZELOG		Logarithm of total assets
INDCOST		A measure of the firm's industry risk
CENTRAL		A measure of the importance of the local managers in choosing the amount of risk to be retained
SOPH		A measure of the degree of importance in using analytical tools

Source: Schmit and Roth (1990).

Table 18

```
Example of the first five observations:
  FIRMCOST ASSUME CAP SIZELOG INDCOST CENTRAL SOPH
1     3.29   0.29   1    9.55    0.32       1   25
2     9.31   0.89   0    8.04    0.33       2   24
3     4.07   1.67   0    7.90    0.34       2   15
4     6.94   1.21   0    8.10    0.34       1   16
5     5.35   0.28   0    7.74    0.09       3   18
```

Table 19. Singapore Automobile Claims

Frees and Valdez (2008) investigated hierarchical models of Singapore driving experience. Here we examine in detail a subset of their data, focusing on 1993 counts of automobile accidents. The purpose of the analysis is to understand the impact of vehicle and driver characteristics on accident experience. These relationships provide a foundation for an actuary working in ratemaking, that is, setting the price of insurance coverages.

The data are from the General Insurance Association of Singapore, an organization consisting of general (property and casualty) insurers in Singapore (see the organization's website: www.gia.org.sg). From this database, several characteristics are available to explain automobile accident frequency. These characteristics include vehicle variables, such as type and age, as well as person level variables, such as age, gender and prior driving experience.

File Name: SingaporeAuto	Number of obs: 7483	Number of variables: 15
Variable	Number of Obs Missing	**Description**
SexInsured		Gender of insured, including male (M), female(F) and unspecified (U)
Female		=1 if female, =0 otherwise
VehicleType		The type of vehicle being insured, such as automobile (A), truck (T), and motorcycle (M)
PC		=1 if private vehicle, =0 otherwise
Clm_Count		Number of claims during the year
Exp_weights		Exposure weight or the fraction of the year that the policy is in effect
LNWEIGHT		Logarithm of exposure weight
NCD		No Claims Discount. This is based ont he previous accident record of the policyholder. The higher the discount, the better is the prior accident record.
AgeCat		The age of the policyholder, in years grouped into seven categories. 0-6 indicate age groups 21 and younger, 22-25, 26-35, 36-45, 46-55, 56-65, 66 and over, respectively
VAgeCat		The age of the vehicle, in years, grouped into seven categories. 0-6 indicate groups 0, 1, 2, 3-5, 6-10, 11-15, 16 and older, respectively
AutoAge0		=1 if private vehicle and VAgeCat = 0, =0 otherwise
AutoAge1		=1 if private vehicle and VAgeCat = 1, =0 otherwise
AutoAge2		=1 if private vehicle and VAgeCat = 2, =0 otherwise
AutoAge		=1 if Private vehicle and VAgeCat = 0, 1 or 2, =0 otherwise
VAgecat1		VAgeCat with categories 0, 1, and 2 combined

Source: Frees and Valdez (2008).

Table 19

Example of the first five observations:

	SexInsured	Female	VehicleType	PC	Clm_Count	Exp_weights	LNWEIGHT	NCD	AgeCat
1	U	0	T	0	0	0.6680356	-0.40341383	30	0
2	U	0	T	0	0	0.5667351	-0.56786326	30	0
3	U	0	T	0	0	0.5037645	-0.68564629	30	0
4	U	0	T	0	0	0.9144422	-0.08944106	20	0
5	U	0	T	0	0	0.5366188	-0.62246739	20	0

	AutoAge0	AutoAge1	AutoAge2	AutoAge	VAgeCat	VAgecat1
1	0	0	0	0	0	2
2	0	0	0	0	0	2
3	0	0	0	0	0	2
4	0	0	0	0	0	2
5	0	0	0	0	0	2

Table 20. Swedish Motor Insurance

These data were compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance, summarized in Hallin and Ingenbleek (1983) and Andrews and Herzberg (1985). The data are cross-sectional, describing third party automobile insurance claims for the year 1977.

The outcomes of interest are the number of claims (the frequency) and sum of payments (the severity), in Swedish kroners. Outcomes are based on 5 categories of distance driven by a vehicle, broken down by 7 geographic zones, 7 categories of recent driver claims experience and 9 types of automobile. Even though there are 2,205 potential distance, zone, experience and type combinations ($5 \times 7 \times 7 \times 9 = 2{,}205$), only n = 2,182 were realized in the 1977 data set.

File Name: SwedishMotorInsurance	Number of obs: 2182	Number of variables: 7
Variable	Number of Obs Missing	**Description**
Kilometres		Distance driven by a vehicle, grouped into five categories
Zone		Graphic zone of a vehicle, grouped into 7 categories
Bonus		Driver claim experience, grouped into 7 categories
Make		The type of a vehicle
Insured		The number of policyholder years. A "policyholder year" is the fraction of the year that the policyholder has a contract with the issuing company.
Claims		Number of claims
Payment		Sum of payments

Source: Hallin and Ingenbleek (1983) and Andrews and Herzberg (1985).

Table 20

```
Example of the first five observations:
  Kilometres Zone Bonus Make Insured Claims Payment
1          1    1     1    1  455.13    108  392491
2          1    1     1    2   69.17     19   46221
3          1    1     1    3   72.88     13   15694
4          1    1     1    4 1292.39    124  422201
5          1    1     1    5  191.01     40  119373
```

Table 21. Term Life Insurance

Like all firms, life insurance companies continually seek new ways to deliver products to the market. Those involved in product development wish to know "who buys insurance and how much do they buy?" Analysts can readily get information on characteristics of current customers through company databases. Potential customers, those that do not have insurance with the company, are often the main focus for expanding market share.

we examine the Survey of Consumer Finances (SCF), a nationally representative sample that contains extensive information on assets, liabilities, in- come, and demographic characteristics of those sampled (potential U.S. customers). We study a random sample of 500 households with positive incomes that were in- terviewed in the 2004 survey.

For term life insurance, the quantity of insurance is measured by the policy FACE, the amount that the company will pay in the event of the death of the named insured. Characteristics that will turn out to be important include annual INCOME, the number of years of EDUCATION of the survey respondent and the number of household members, NUMHH.

File Name: TermLife	Number of obs: 500	Number of variables: 18
Variable	Number of Obs Missing	**Description**
GENDER		Gender of the survey respondent
AGE		Age of the survey respondent
MARSTAT		Marital status of the survey respondent (=1 if married, =2 if living with partner, and =0 otherwise)
EDUCATION		Number of years of education of the survey respondent
ETHNICITY		Ethnicity
SMARSTAT		Marital status of the respondent's spouse
SGENDER		Gender of the respondent's spouse
SAGE		Age of the respondent's spouse
SEDUCATION		Education of the respondent's spouse
NUMHH		Number of household members
INCOME		Annual income of the family
TOTINCOME		Total income
CHARITY		Charitable contributions
FACE		Amount that the company will pay in the event of the death of the named insured
FACECVLIFEPOLICIES		Face amount of life insurance policy with a cash value
CASHCVLIFEPOLICIES		Cash value of life insurance policy with a cash value
BORROWCVLIFEPOL		Amount borrowed on life insurance policy with a cash value
NETVALUE		Net amount at risk on life insurance policy with a cash value

Source: Survey of Consumer Finances (SCF).

Table 21

Example of the first five observations:

	GENDER	AGE	MARSTAT	EDUCATION	ETHNICITY	SMARSTAT	SGENDER	SAGE	SEDUCATION	NUMHH	INCOME
1	1	30	1	16	3	2	2	27	16	3	43000
2	1	50	1	9	3	1	2	47	8	3	12000
3	1	39	1	16	1	2	2	38	16	5	120000
4	1	43	1	17	1	1	2	35	14	4	40000
5	1	61	1	15	1	2	2	59	12	2	25000

	TOTINCOME	CHARITY	FACE	FACECVLIFEPOLICIES	CASHCVLIFEPOLICIES	BORROWCVLIFEPOL
1	43000	0	20000	0	0	0
2	0	0	130000	0	0	0
3	90000	500	1500000	0	0	0
4	40000	0	50000	75000	0	5
5	1020000	500	0	7000000	300000	5

	NETVALUE
1	0
2	0
3	0
4	0
5	0

Table 22. National Life Expectancies

Who is doing health care right? Health care decisions are made at the individual, corporate and government levels. Virtually every person, corporation and government have their own perspective on health care; these different perspectives result in a wide variety of systems for managing health care. Comparing different health care systems help us learn about approaches other than our own, which in turn help us make better decisions in designing improved systems.

Here, we consider health care systems from $n = 185$ countries throughout the world. As a measure of the quality of care, we use LIFEEXP, the life expectancy at birth. There are 185 countries consider in this study, not all countries provided information for each variable. Data not available are noted under the column "Number Missing".

File Name: UNLifeExpectancy	Number of obs: 185	Number of variables: 15
Variable	Number of Obs Missing	**Description**
REGION		Categorical variable for region of the world
COUNTRY		The name of the country
LIFEEXP		Life expectancy at birth, in years
ILLITERATE	14	Adult illiteracy rate, % aged 15 and older
POP	1	2005 population, in millions
FERTILITY	4	Total fertility rate, births per woman
PRIVATEHEALTH	1	2004 Private expenditure on health, % of GDP
PUBLICEDUCATION	28	Public expenditure on education, % of GDP
HEALTHEXPEND	5	2004 Health expenditure per capita, PPP in USD
BIRTHATTEND	7	Births attended by skilled health personnel (%)
PHYSICIAN	3	Physicians per 100,000 people
SMOKING	88	Prevalence of smoking, (male) % of adults
RESEARCHERS	95	Researchers in R & D, per million people
GDP	7	Gross domestic product, in billions of USD
FEMALEBOSS	87	Legislators, senior officials and managers, % female

Source: United Nations Human Development Report, available at http://hdr.undp.org/en/.

Table 22

Example of the first five observations:

	REGION	COUNTRY	LIFEEXP	ILLITERATE	POP	FERTILITY	PRIVATEHEALTH
1	4	Afghanistan	42.9	72.0	25.1	7.5	3.7
2	7	Albania	76.2	1.3	3.2	2.2	3.7
3	1	Algeria	71.7	30.1	32.9	2.5	1.0
4	6	Angola	41.7	32.6	16.1	6.8	0.4
5	3	Antigua and Barbuda	73.9	14.2	0.1	NA	1.4

	PUBLICEDUCATION	HEALTHEXPEND	BIRTHATTEND	PHYSICIAN	SMOKING	RESEARCHERS	GDP
1	NA	19	14	19	NA	NA	7.3
2	2.9	339	98	131	60	NA	8.4
3	NA	167	96	113	32	NA	102.3
4	2.6	38	45	8	NA	NA	32.8
5	3.8	516	100	17	NA	NA	0.9

	FEMALEBOSS
1	NA
2	NA
3	NA
4	NA
5	45

Table 23. Nursing Home Utilization

The nursing home data are provided by the Wisconsin Department of Health and Family Services (DHFS). The State of Wisconsin Medicaid program funds nursing home care for individuals qualifying on the basis of need and financial status. As part of the conditions for participation, Medicaid-certified nursing homes must file an annual cost report to 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. The data are publicly available; see http://dhfs.wisconsin.gov/provider/prev-yrs-reports-nh.htm for more information.

The DHFS is interested in predictive techniques that provide reliable utilization forecasts to update their Medicaid funding rate schedule of nursing facilities. The data here is in cost report years 2000 and 2001. There are 362 facilities in 2000 and 355 facilities in 2001. Typically, utilization of nursing home care is measured in patient days ("patient days" is the number of days each patient was in the facility, summed over all patients).

File Name: WiscNursingHome	Number of obs: 717	Number of variables: 12
Variable	Number of Obs Missing	**Description**
hospID		Hospital identification number
CRYEAR		Cost report year
TPY		Total patient years
NUMBED		Number of beds
SQRFOOT	10	Square footage of the nursing home
MSA		Metropolitan Statistical Area code, 1-13, 0 for rural
URBAN		1 if urban, 0 if rural
PRO		1 if for profit, 0 for non-profit
TAXEXEMPT		1 if tax-exempt
SELFFUNDINS		1 if self-funded for insurance
MCERT		1 if Medicare certified
ORGSTR		1 for profit, 2 for tax-exempt, 3 for governmental unit

Source: Rosenberg et al. (2008).

Table 23

Example of the first five observations:

	ID	CRYEAR	TPY	NUMBED	SQRFOOT	MSA	URBAN	PRO	TAXEXEMPT	SELFFUNDINS	MCERT	ORGSTR
1	101	2000	16.48087	18	10.861	0	0	0	1	0	0	2
2	103	2000	59.24590	63	19.782	0	0	0	0	1	1	3
3	105	2000	49.63661	54	26.868	1	1	1	0	1	1	1
4	107	2000	51.87432	60	26.319	0	0	0	1	1	1	2
5	108	2000	94.56011	104	30.700	10	1	1	0	0	1	1

Table 24. Wisconsin Hospital Costs

Identifying predictors of hospital charges can provide direction for hospitals, government, insurers and consumers in controlling these factors that in turn leads to better control of hospital costs. We study the impact of various predictors on hospital charges in the state of Wisconsin. The data for the year 1989 were obtained from the Office of Health Care Information, Wisconsin's Department of Health and Human Services. Cross sectional data are used, which details the 20 diagnosis related group (DRG) discharge costs for hospitals in the state of Wisconsin, broken down into nine major health service areas and three types of payer (Fee for service, HMO, and other). Even though there are 540 potential DRG, area and payer combinations ($20 \times 9 \times 3 = 540$), only 526 combinations were actually realized in the 1989 data set. Other predictor variables included the logarithm of the total number of discharges (NO DSCHG) and total number of hospital beds (NUM BEDS) for each combination. The response variable is the logarithm of total hospital charges per number of discharges (CHGNUM).

File Name: WiscHospCosts	Number of obs: 526	Number of variables: 9
Variable	Number of Obs Missing	**Description**
TOT_CHG		Hospital discharged costs
HSA		Health service area A variable that categorizes Wisconsin into nine areas
DRG		Diagnostic related group, a classification code to label the reason for hospital care
PAYER		The type of payer, 1 if fee-for-service, 2 if health maintenance organization, and 3 otherwise
NO_DSCHG		Number of patients discharged from the hospital
POPLN		Size of the area population
NUM_EDS		Number of hospital beds, a measure of capacity
INCOME		Average income within the area, a measure of the ability to pay for hospital utilization
CHG_NUM		Hospital discharged costs per patient

Source: Office of Health Care Information, Wisconsin Department of Health and Human Services.

Table 24

Example of the first five observations:

	TOT_CHG	HSA	DRG	PAYER	NO_DSCHG	POPLN	NUM_BEDS	INCOME	CHG_NUM
1	5810558	1	14	1	1164	869000	3256	10355	4991.888
2	463455	1	14	2	65	869000	3256	10355	7130.077
3	585057	1	14	3	91	869000	3256	10355	6429.198
4	5004093	1	89	1	1084	869000	3256	10355	4616.322
5	254151	1	89	2	60	869000	3256	10355	4235.850

Table 25. Wisconsin Lottery Sales

State of Wisconsin lottery administrators are interested in assessing factors that affect lottery sales. This data set described a sample of 50 geographic areas (zip codes) containing sales data on the Wisconsin state lottery (SALES). Sales consists of online lottery tickets that are sold by selected retail establishments in Wisconsin. These tickets are generally priced at $1.00, so the number of tickets sold equals the lottery revenue. We analyze average lottery sales (SALES) over a forty-week period, April, 1998 through January, 1999, from fifty randomly selected areas identified by postal (ZIP) code within the state of Wisconsin.

File Name: WiscLottery	Number of obs: 50	Number of variables: 10
Variable	Number of Obs Missing	**Description**
ZIP		Zip code within the state of Wisconsin
PERPERHH		Persons per household
MEDSCHYR		Median years of schooling
MEDHVL		Median home value in $1000s for owner-occupied homes
PRCRENT		Percent of housing that is renter-occupied
PRC55P		Percent of population that is 55 or older
HHMEDAGE		Household median age
MEDINC		Estimated median household income, in $1000s
SALE		Online lottery sales to individual consumers
POP		Population, in thousands

Source: Frees and Miller (2003).

Table 25

Example of the first five observations:

	ZIP	PERPERHH	MEDSCHYR	MEDHVL	PRCRENT	PRC55P	HHMEDAGE	MEDINC	SALES	POP
1	53003	3.0	12.6	71.3	21	38	48	54.2	1285.400	435
2	53033	3.2	12.9	98.0	6	28	46	70.7	3571.450	4823
3	53038	2.8	12.4	58.7	25	35	45	43.6	2407.037	2469
4	53059	3.1	12.5	65.7	24	29	45	51.9	1223.825	2051
5	53072	2.6	13.1	96.7	32	27	42	63.1	15046.400	13337

Table 26. Workers Compensation

We consider a standard example in worker's compensation insurance, examining losses due to permanent, partial disability claims. The data are from Klugman (1992), who considers Bayesian model representations, and are originally from the National Council on Compensation Insurance. We consider n=121 occupation, or risk, classes, over T=7 years. To protect the data source, further information on the occupation classes and years is not available. Source: Frees, E. W., Young, V. and Y. Luo (2001). Case studies using panel data models. North American Actuarial Journal, 4, No. 4, 24-42.

File Name: WiscLottery	Number of obs: 847	Number of variables: 4
Variable	Number of Obs Missing	**Description**
CL		Occupation class identifier, 1-124
YR		Year identifier, 1-4
PR		Payroll, a measure of exposure to loss, in tens of millions of dollars
LOSS		Losses related to permanent partial disability, in tens of millions of dollars

Source: Klugman (1992).

Table 26

```
Example of the first five observations:
CL YR       PR    LOSS
1   1 21798086  538707
1   2 22640528  439184
1   3 22572010 1059775
1   4 24789710  560013
1   5 25876764 1004997
```

Table 27. Euro Exchange Rates

The exchange rate that we consider is the amount of Euros that one can purchase for one U.S. dollar. We have $T = 699$ daily observations from the period April 1, 2005 through January 8, 2008. These data were obtained from the Federal Reserve (H10 report).

Note: The data are based on noon buying rates in New York from a sample of market participants and they represent rates set for cable transfers payable in the listed currencies. These are also the exchange rates required by the Securities and Exchange Commission for the integrated disclosure system for foreign private issuers.

File Name: EuroExchange	Number of obs: 699	Number of variables: 3
Variable	Number of Obs Missing	**Description**
date		Calendar date
exhkus		The number of Hong Kong dollars that one can purchase for one U.S. dollar
exeuus		The number of Euro dollars that one can purchase for one U.S. dollar

Source: Federal Reserve Bank of New York.

Table 27

```
Example of the first five observations:
      date exhkus exeuus
1 04/01/05 7.7989 0.7754
2 04/04/05 7.7991 0.7789
3 04/05/05 7.7995 0.7787
4 04/06/05 7.7993 0.7771
5 04/07/05 7.7990 0.7748
```

Table 28. Hong Kong Exchange Rates

For travelers and firms, exchange rates are an important part of the monetary economy. The exchange rate that we consider here is the number of Hong Kong dollars that one can purchase for one U.S. dollar. We have $T = 502$ daily observations for the period April 1, 2005 through May 31, 2007 that were obtained from the Federal Reserve (H10 report).

File Name: HKExchange	Number of obs: 502	Number of variables: 3
Variable	Number of Obs Missing	**Description**
DATE		Calendar date
EXHKUS		The number of Hong Kong dollars that one can purchase for one U.S. dollar
EXEUROUS		The number of Euro dollars that one can purchase for one U.S. dollar

Source: Foreign Exchange Rates (Federal Reserve, H10 report).

Table 28

```
Example of the first five observations:
      DATE EXHKUS EXEUROUS
1 1-Apr-05 7.7989   0.7754
2 4-Apr-05 7.7991   0.7789
3 5-Apr-05 7.7995   0.7787
4 6-Apr-05 7.7993   0.7771
5 7-Apr-05 7.7990   0.7748
```

Table 29. Inflation Bond Prices

Beginning in January of 2003, the US Treasury Department established an inflation bond index that summarizes the returns on long-term bonds offered by the Treasury Department that are inflation-indexed. For a treasury inflation protected security (TIPS), the principal of the bond is indexed by the (three month lagged) value of the (non-seasonally adjusted) consumer price index. The bond then pays a semi-annual coupon at a rate determined at auction when the bond is issued. The index that we examine is the unweighted average of bid yields for all TIPS with remaining terms to maturity of 10 or more years (Source: *US Treasury*). Monthly values of the index from January 2003 through March 2007 are considered, for a total of $T = 51$ returns.

File Name: InflationBond	Number of obs: 51	Number of variables: 2
Variable	Number of Obs Missing	**Description**
date		Calendar date
INFBOND		Inflation Bond Index that summarizes the returns on long-term bonds offered by the Treasury Department that are inflation-indexed

Source: US Treasury.

Table 29

```
Example of the first five observations:
       date INFBOND
1 31-Jan-03    2.72
2 28-Feb-03    2.50
3 31-Mar-03    2.52
4 30-Apr-03    2.72
5 31-May-03    2.40
```

Table 30. Labor Force Participation Rate

Labor force participation rate ($LFPR$) forecasts, coupled with forecasts of the population, provide us with a picture of a nation's future workforce. This picture provides insights to the future workings of the overall economy, and thus $LFPR$ projections are of interest to a number of government agencies. In the United States, $LFPR$s are projected by the Social Security Administration, the Bureau of Labor Statistics, the Congressional Budget Office and the Office of Management and Budget. In the context of Social Security, policy-makers use labor force projections to evaluate proposals for reforming the Social Security system and to assess its future financial solvency.

The labor force participation rates are the civilian labor force divided by the civilian non-institutional population. These data are compiled by the Bureau of Labor Statistics. For illustration purposes, we examine a specific demographic cell and show how to forecast it - forecasts of other cells may be found in Fullerton (1999) and Frees (2006). Specifically, we examine 1968-1998 for females, aged 20-44, living in a household with a spouse present and at least one child under six years of age.

File Name: LaborForcePR	Number of obs: 31	Number of variables: 3
Variable	Number of Obs Missing	**Description**
TIME		Time, 1, ..., 31
YEAR		Calendar year
MSC6U		Labor Force Participation Rates for females aged 20-24, living in a household with a spouse present and at least one child under six years of age

Source: Census Bureau.

Table 30

```
Example of the first five observations:
  TIME YEAR     MSC6U
1    1 1968 0.2778812
2    2 1969 0.2868593
3    3 1970 0.3065709
4    4 1971 0.2981785
5    5 1972 0.3053743
```

Table 31. Medical Component of the CPI

The CPI is a breadbasket of goods and services whose price is measured by the Bureau of Labor Statistics. By measuring this breadbasket periodically, consumers get an idea of the steady increase in prices over time which, among other things, serves as a proxy for inflation. The CPI itself is composed of many components, reflecting the relative importance of each component to the overall economy. Here, we study the medical component of the CPI, the fastest growing part of the overall breadbasket since 1967. The data we consider are quarterly values of the medical component of the CPI (MCPI) over a sixty year period from 1947 to the first quarter of 2007, inclusive. Over this period, the index rose from 13.3 to 346.0. This represents a twenty-six fold increase over the sixty year period which translates roughly into a 1.36% quarterly increase.

File Name: MedCPISmooth	Number of obs: 241	Number of variables: 10
Variable	Number of Obs Missing	**Description**
yearInt		Calendar year
Month		The last month of the quarter
Quarter		Number of quarter
value		Quarterly values of the medical component of the CPI (MCPI)
PerMEDCPI	1	Quarterly increase of the medical component of the CPI, in percent
YEAR		Year plus a fraction for the quarter
MCPISM4	1	Medical component of consumer's price index, smoothed with k=4
MCPISM8	1	Medical component of consumer's price index, smoothed with k=8
MCPISMw_2	1	Medical component of consumer's price index, smoothed with w=.2
MCPISMw_8	1	Medical component of consumer's price index, smoothed with w=.8

Source: Bureau of Labor Statistics.

Table 31

Example of the first five observations:

	yearInt	Month	Quarter	value	PerMEDCPI	YEAR	MCPISM4	MCPISM8	MCPISMw_2	MCPISMw_8
1	1947	3	1	13.3	NA	1947.167	NA	NA	NA	NA
2	1947	6	1	13.5	1.504	1947.417	1.504	1.504	1.504	1.504
3	1947	9	1	13.7	1.481	1947.667	1.493	1.493	1.486	1.499
4	1947	12	1	13.9	1.460	1947.917	1.482	1.482	1.465	1.491
5	1948	3	1	14.1	1.439	1948.167	1.471	1.471	1.444	1.481

Table 32. Medicare Hospital Costs

We consider T=6 years, 1990-1995, of data for inpatient hospital charges that are covered by the Medicare program. The data were obtained from the Health Care Financing Administration, Bureau of Data Management and Strategy. To illustrate, in 1995 the total covered charges were 157.8 billions for twelve million discharges. For this analysis, we use state as the subject, or risk class. Thus, we consider n=54 states that include the 50 states in the Union, the District of Columbia, Virgin Islands, Puerto Rico and an unspecified "other" category.

File Name: Medicare	Number of obs: 324	Number of variables: 9
Variable	Number of Obs Missing	**Description**
STATE		State identifier, 1-54
YEAR		Year identifier, 1-6
TOT_CHG		Total hospital charges, in millions of dollars.
COV_CHG		Total hospital charges covered by Medicare, in millions of dollars.
MED_REIM		Total hospital charges reimbursed by the Medicare program, in millions of dollars.
TOT_D		Total number of hospitals stays, in days.
NUM_DSHG		Number discharged, in thousands.
AVE_T_D		Average hospital stay per discharge in days.
NMSTATE		Name of the state.

Source: Frees, Young, and Luo (2001)

Table 32

Example of the first five observations:

	STATE	YEAR	TOT_CHG	COV_CHG	MED_REIB	TOT_D	NUM_DCHG	AVE_T_D	NMSTATE
1	1	1	2211617271	2170240349	972752944	1932673	230015	8	AL
2	1	2	2523987347	2468263759	1046016144	1936939	234739	8	AL
3	1	3	2975969979	2922611694	1205791592	2016354	245027	8	AL
4	1	4	3194595003	3149745611	1307982985	1948427	243947	8	AL
5	1	5	3417704863	3384305357	1376211788	1926335	258384	7	AL

Table 33. Prescription Drug Prices

We consider a series from the State of New Jersey's Prescription Drug Program, the cost per prescription claim. This monthly series is available over the period August, 1986 through March, 1992, inclusive. It shows that the series is clearly nonstationary, in that cost per prescription claims are increasing over time. There are a variety of ways of handling this trend. One may begin with a linear trend in time and include lag claims to handle autocorrelations. For this series, a good approach to the modeling turns out to be to consider the percentage changes in the cost per claim series.

File Name: PrescriptionDrug	Number of obs: 68	Number of variables: 10
Variable	Number of Obs Missing	**Description**
PAID_CLM		Monthly paid prescription claims paid by the New Jersey Prescription Drug Program
NPRESCRP		Number of prescription claims
TIME		Sequence number, time
MONTH		Month
COST_CLM		Cost per claim, defined to be (PAID_CLM)/NPRESCRP
RATEC_C	1	Rate of change of the cost per claim as a percentage, defined by 100*[(current COST_CLM/previous COST_CLM)-1]
SINET		The sine function evaluated at TIME
COST		The cosine function evaluated at TIME
SINE2T		The sine function evaluated at 2*TIME
COS2T		The cosine function evaluated at 2*TIME

Source: Frees (1995).

Table 33

Example of the first five observations:

	PAID_CLM	NPRESCRP	TIME	MONTH	COST_CLM	RATEC_C	SINET	COST
1	992300	68213	1	8	14.54708	NA	0.5000000	8.660254e-01
2	1143249	77920	2	9	14.67209	0.8593202	0.8660254	5.000000e-01
3	935150	63179	3	10	14.80160	0.8826852	1.0000000	-4.371140e-08
4	962309	65855	4	11	14.61254	-1.2772441	0.8660254	-5.000001e-01
5	1106053	77364	5	12	14.29674	-2.1611750	0.5000001	-8.660254e-01

	SINE2T	COS2T
1	8.660254e-01	0.5000000
2	8.660254e-01	-0.5000001
3	-8.742280e-08	-1.0000000
4	-8.660254e-01	-0.4999999
5	-8.660254e-01	0.4999999

Table 34. Standard and Poor's 500 Daily

These data consists of the 1759 daily returns for the calendar years 2000 through 2006 of the Standard and Poor's (S&P) value weighted index. Each year, there are about 250 days on which the exchange is open and stocks were traded - on weekends and holidays it is closed. For each trading day an average of the closing, or last, price of various stocks were taken to form the S&P equally weighted index for that day. There are several indices to measure the market's overall performance. The value weighted index is created by assuming that the amount invested in each stock is proportional to its market capitalization. Here, the market capitalization is simply the beginning price per share times the number of outstanding shares.

Financial economic theory states that if the market were predictable, many investors would attempt to take advantage of these predictions, thus forcing unpredictability. For example, suppose a statistical model reliably predicted mutual fund A to increase two-fold over the next 18 months. Then, the no arbitrage principle in financial economics states that several alert investors, armed with information from the statistical model, would bid to buy mutual fund A, thus causing the price to increase because demand is increasing. These alert investors would continue to purchase until the price of mutual fund A rose to the point where the return was equivalent to other investment opportunities in the same risk class. Any advantages produced by the statistical model would disappear rapidly, thus eliminating this advantage.

Thus, financial economic theory states that for liquid markets such as stocks represented through the S&P index there should be no detectable patterns, resulting in a white noise process. In practice, it has been found that cost of buying and selling equities (called transactions costs) are large enough so as to prevent us from taking advantage of these slight tendencies in the swings of the market. This illustrates a point known as statistically significant but not practically important. This is not to suggest that statistics is not practical (heavens forbid!). Instead, statistics in and of itself does not explicitly recognize factors, such as economic, psychological and so on, that may be extremely important in any given situation. It is up to the analyst to interpret the statistical analysis in light of these factors.

File Name: SP500Daily	Number of obs: 1759	Number of variables: 2
Variable	Number of Obs Missing	**Description**
caldt		Calendar date
vwretd		The Standard and Poor's 500 daily value weighted return

Source: Center for Research on Security Prices, University of Chicago.

Table 34

```
Example of the first five observations:
     caldt        vwretd
1 20000103 -0.0093845450
2 20000104 -0.0384355500
3 20000105  0.0008613558
4 20000106 -0.0028339380
5 20000107  0.0321512800
```

Table 35. Standard and Poor's 500 Quarterly

An important task of a financial analyst is to quantify costs associated with future cash flows. We consider here funds invested in a standard measure of overall market performance, the Standard and Poor's (S&P) 500 Composite Index. The goal is to forecast the performance of the portfolio for discounting of cash flows. In particular, we examine the S&P Composite Quarterly Index for the years 1936 to 2007, inclusive.

File Name: SP500Quarterly	Number of obs: 284	Number of variables: 5
Variable	Number of Obs Missing	**Description**
YEAR		Year
SPINDEX		The Standard and Poor's (S&P) 500 Composite Index
DIFFINDEX		The difference of the SPINDEX between this year and last year
LNSPINDEX		The natural logarithm of SPINDEX
DIFFLNSP		The difference of LNSPINDEX between this year and last year

Source: Center for Research on Security Prices, University of Chicago.

Table 35

Example of the first five observations:

	YEAR	SPINDEX	DIFFINDEX	LNSPINDEX	DIFFLNSP
1	1936.166667	14.92000008	0	2.7027026	0
2	1936.416667	14.84000015	-0.079999923	2.697326248	-0.005376352
3	1936.666667	16.01000023	1.170000076	2.773213541	0.075887293
4	1936.916667	17.18000031	1.170000076	2.843745934	0.070532393
5	1937.166667	17.92000008	0.739999771	2.885917412	0.042171477

Table 36. Auto Industry

The data represent industry aggregates for private passenger auto liability/medical coverages from year 2004, in millions of dollars. They are based on insurance company annual statements, specifically, Schedule P, Part 3B. The elements of the triangle represent cumulative net payments, including defense and cost containment expenses.

File Name: IndustryAuto	Number of obs: 55	Number of variables: 3
Variable	Number of Obs Missing	**Description**
Incurral Year		The year in which a claim has been incurred
Development Year		The number of years from incurral to the time when the payment is made
Claim		Cumulative net payments, including defense and cost containment expenses

Source: Wacek (2007).

Table 36

```
Example of the first five observations:
    Incurral Year Development Year Claim
1            1995                1 17674
2            1996                1 18315
3            1997                1 18606
4            1998                1 18816
5            1999                1 20649
```

Table 37. Medical Care

These data for 36 months of medical care payments, from January 2001 through December 2003, inclusive. These are payments for medical care coverage with no deductible nor coinsurance. There were relatively low co-payments, such as $10 per visit. The payments exclude prescription drugs that typically have a shorter payment pattern compared with other medical claims.

File Name: MedicalCare	Number of obs: 390	Number of variables: 4
Variable	Number of Obs Missing	**Description**
Members		Total number of members
Month		The month in which a claim has been incurred
Delay		The number of months from incurral to the time when the payment is made
Payments		The payments excluding prescription drugs that typically have a shorter payment pattern compared with other medical claims

Source: Gamage et al. (2007).

Table 37

```
Example of the first five observations:
  Members Month Delay Payments
1   11154     1     1      180
2   11118     2     1     5162
3   11070     3     1    42263
4    1106     4     1    20781
5   11130     5     1    20346
```

Table 38. Reinsurance General Liability

The data originate from the 1991 edition of the "Historical Loss Development Study" published by the Reinsurance Association of American (page 91). These data have been widely used to illustrate triangle methods, beginning with Mack (1994) and later by England and Verrall (2002). These data are from automatic facultative reinsurance business in general liability (excluding asbestos and environmental) coverages. (Under a facultative basis, each risk is underwritten by the reinsurer on its own merits.) The data contains data for years 1981-1990, inclusive.

File Name: ReinsGenLiab	Number of obs: 55	Number of variables: 3
Variable	Number of Obs Missing	**Description**
Incurral Year		The year in which a claim has been incurred
Development Year		The number of years from incurral to the time when the payment is made
Claim		Incremental incurred losses in thousands of US dollars

Source: The Reinsurance Association of American.

Table 38

```
Example of the first five observations:
  Incurral Year Development Year Claim
1             1                1  5012
2             2                1   106
3             3                1  3410
4             4                1  5655
5             5                1  1092
```

Table 39. Reinsurance General Liability 2004

The data is an excerpt from Braun (2004) that is based on the 2001 edition "Historical Loss Development Study" published by the Reinsurance Association of American. The data contains data for years 1987-2000, inclusive.

File Name: ReinsGL2004	Number of obs: 105	Number of variables: 3
Variable	Number of Obs Missing	**Description**
Incurral Year		The year in which a claim has been incurred
Development Year		The number of years from incurral to the time when the payment is made
Claim		Incremental incurred losses from 1995-2000, in thousands of US dollars

Source: The Reinsurance Association of American.

Table 39

```
Example of the first five observations:
  Incurral Year DevelopmentYear Claim
1          1987               1 59966
2          1988               1 49685
3          1989               1 51914
4          1990               1 84937
5          1991               1 98921
```

Table 40. Singapore Auto Injury

The data contain payments from a portfolio of automobile policies for a Singapore property and casualty (general) insurer. Payments, deflated for inflation, are for third party injury from comprehensive insurance policies. The data are for policies with coverages from 1993-2001, inclusive.

File Name: SingaporeInjury	Number of obs: 45	Number of variables: 3
Variable	Number of Obs Missing	**Description**
Month		The month in which a claim has been incurred
Delay		The number of months from incurral to the time when the payment is made
Payments		Incremental payments, deflated for inflation, for third party injury from comprehensive insurance policies.

Source: Frees and Valdez (2008).

Table 40

```
Example of the first five observations:
  Year Delay   Payment
1 1993     1  14695
2 1994     1 153615.21
3 1995     1  24741.26
4 1996     1  68630.4
5 1997     1  29177.17
```

Table 41. Singapore Auto Property Damage

The data report incremental payments from a portfolio of automobile policies for a Singapore property and casualty (general) insurer. Here, payments are for third party property damage from comprehensive insurance policies. All payments have been deflated using a Singaporean consumer price index, so they are in constant dollars. The data are for policies with coverages from 1997-2001, inclusive.

File Name: SingaporeProperty	Number of obs: 15	Number of variables: 3
Variable	Number of Obs Missing	**Description**
Year		The year in which a claim has been incurred
Delay		The number of years from incurral to the time when the payment is made
Payments		Incremental payments, deflated for inflation, for third party property damage from comprehensive insurance policies.

Source: Frees and Valdez (2008).

Table 41

```
Example of the first five observations:
  Year Delay   Payments
1 1997     1 1188675
2 1998     1 1235401.82
3 1999     1 2209849.65
4 2000     1 2662545.97
5 2001     1 2457265.33
```