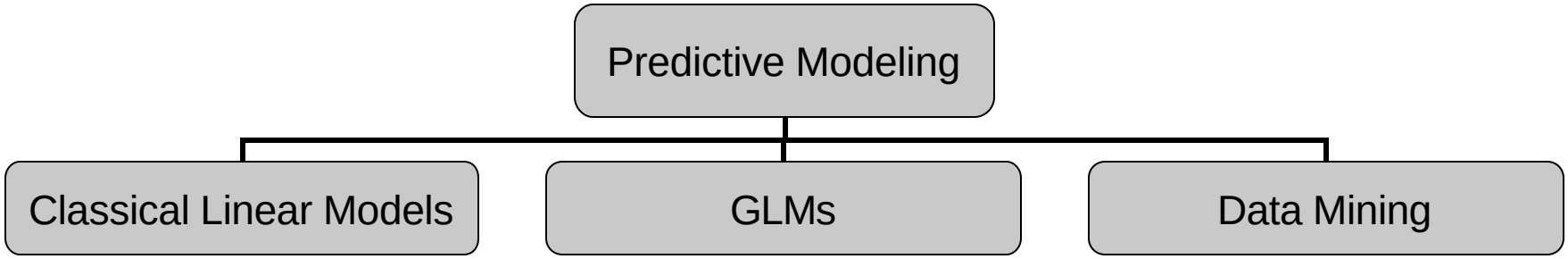
Introduction to Predictive Modeling

Prepared by
Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
June 20, 2005

Modeling 101 Objectives

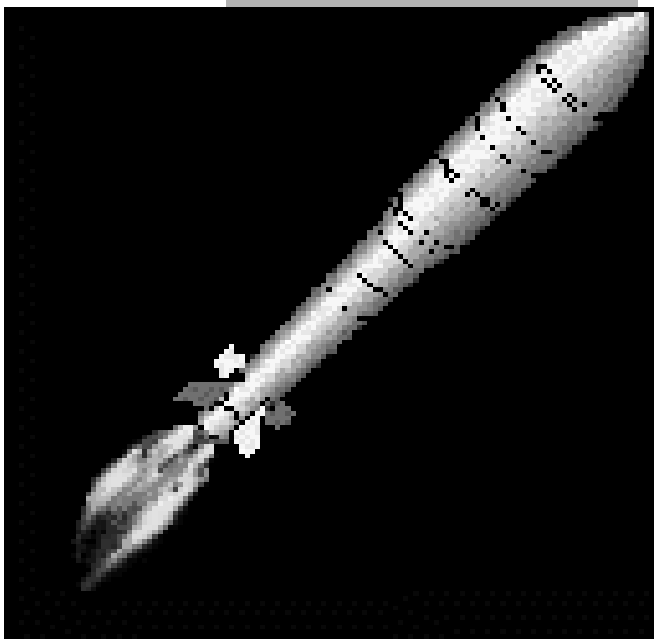
- Gentle introduction to classical statistical models and
- Introduction to some more advanced models
- Illustrate some simple applications
- Show examples in commonly available software (see Excel files that accompany slides)
- Discuss practical modeling issues
- Which model(s) to use?

Predictive Modeling Family



Why Predictive Modeling?

- Better use of data than traditional methods
- Advanced methods for dealing with messy data now available



Major Kinds of Modeling

- Supervised learning
  - Most common situation
  - A dependent variable
    - Frequency
    - Loss ratio
    - Fraud/no fraud
  - Some methods
    - Regression
    - CART
    - Some neural networks
- Unsupervised learning
  - No dependent variable
  - Group like records together
    - A group of claims with similar characteristics might be more likely to be fraudulent
  - Some methods
    - Association rules
    - K-means clustering
    - Kohonen neural networks

Kinds of Applications

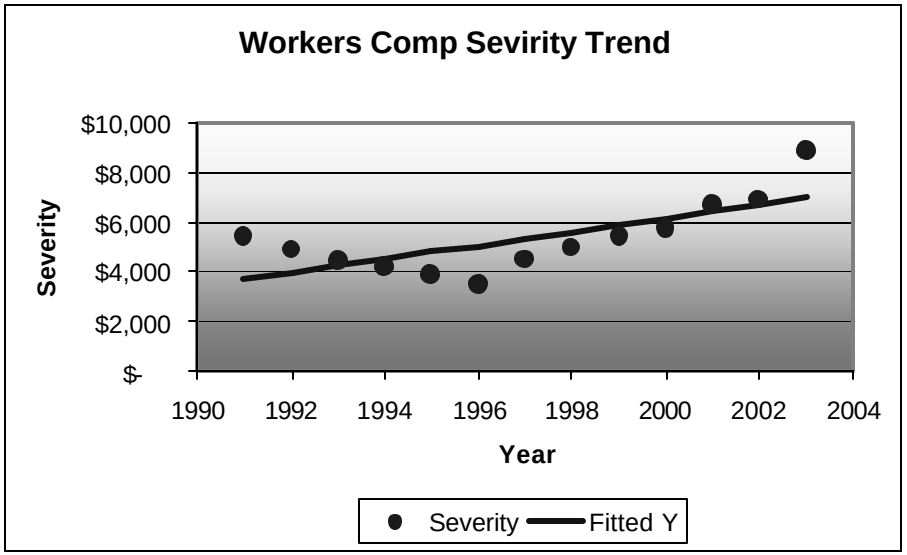
- Classification
- Prediction

Common Example: Fit Severity Trend

- Example data and analyses in Trend Projection Pred Modl.xls

A Brief Introduction to Regression for Prediction

- One of most common statistical methods fits a line to data
- Model: Y = a+bx + error
- Error assumed to be Normal



A Brief Introduction to Regression

- Fits line that minimizes squared deviation between actual and fitted values
- $$\min\left(\sum (Y_i - \widehat{Y})^2\right)$$

$$b = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}, \quad a = \bar{Y} - b\,\bar{X}$$

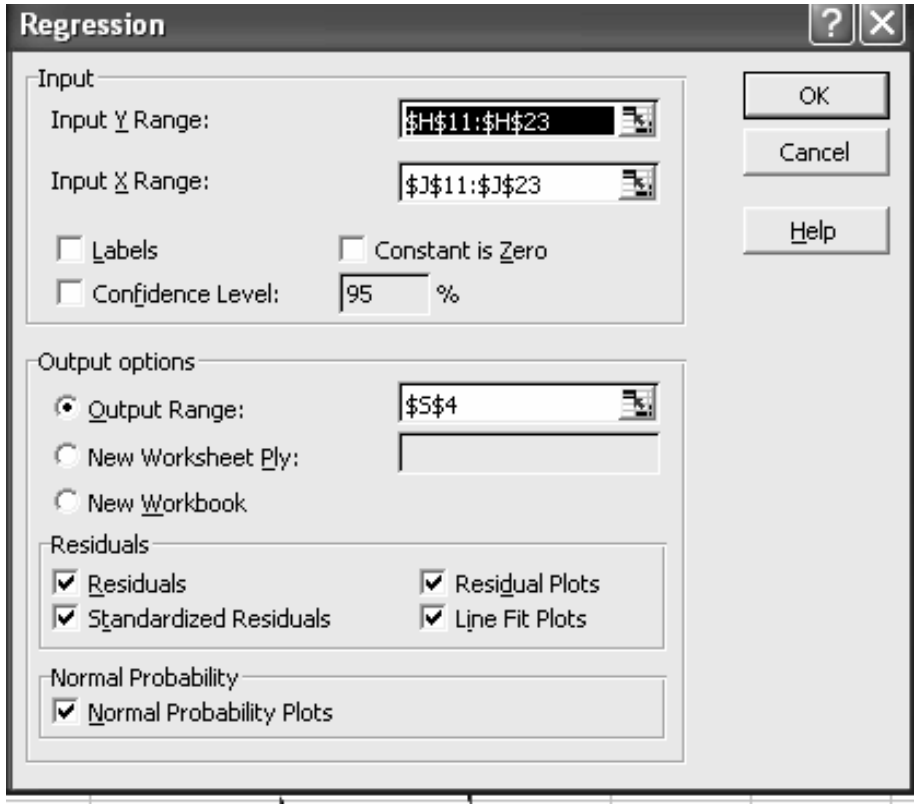
Simple Formula for Fitting Line

$$b = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}, \quad a = \bar{Y} - b\,\bar{X}$$

(1) Accident Year	(2) Severity Y	(3) Fitted Y	(4) X	(5) $Y - \bar{Y}$	(6) $X - \bar{X}$	(7) $(Y - \bar{Y})(X - \bar{X})$	(8) $(X - \bar{X})^2$
1991	$ 5,410	$ 3,715	1	$ 75	-6	(448.91)	36
1992	$ 4,868	$ 3,985	2	$ (467)	-5	2,336.63	25
1993	$ 4,393	$ 4,255	3	$ (943)	-4	3,770.32	16
1994	$ 4,191	$ 4,525	4	$ (1,145)	-3	3,434.25	9
1995	$ 3,892	$ 4,796	5	$ (1,443)	-2	2,886.80	4
1996	$ 3,494	$ 5,066	6	$ (1,842)	-1	1,841.86	1
1997	$ 4,529	$ 5,336	7	$ (806)	0	-	0
1998	$ 4,977	$ 5,606	8	$ (358)	1	(358.10)	1
1999	$ 5,453	$ 5,876	9	$ 117	2	234.45	4
2000	$ 5,727	$ 6,146	10	$ 391	3	1,174.07	9
2001	$ 6,687	$ 6,416	11	$ 1,351	4	5,405.06	16
2002	$ 6,885	$ 6,686	12	$ 1,550	5	7,747.87	25
2003	$ 8,855	$ 6,956	13	$ 3,520	6	21,119.29	36
Sum/Average	$ 5,336		7			49,143.61	182.00
B						270.02	
$a = \bar{Y} - b\,\bar{X}$						3,445.41	

Excel Does Regression

- Install Data Analysis Tool Pak (Add In) that comes with Excel
- Click Tools, Data Analysis, Regression



Key Statistic: Residual

The Residual plays a key role in regression evaluation and diagnostics

$$\text{Re}\,sidual_i = Y_i - \widehat{Y}_i$$

$\widehat{Y}_i$ is model estimate for Y_i

$SSR =$ Sum Squared Residual

$$= \sum_{1}^{n} (Y_i - \widehat{Y}_i)^2$$

Goodness of Fit Statistics

- R^2: (SS Regression/SS Total)
  - percentage of variance explained
- F statistic: (MS Regression/MS Resid)
  - significance of regression
- T statistics: Uses SE of coefficient to determine if it is significant
  - significance of coefficients
  - It is customary to drop variable if coefficient not significant

    - Note SS = Sum squared of errors

Output of Excel Regression Procedure

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.72
R Square	0.52
Adjusted R Square	0.48
Standard Error	1052.73
Observations	13.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	13269748.70	13269748.70	11.97	0.01
Residual	11	12190626.36	1108238.76		
Total	12	25460375.05			

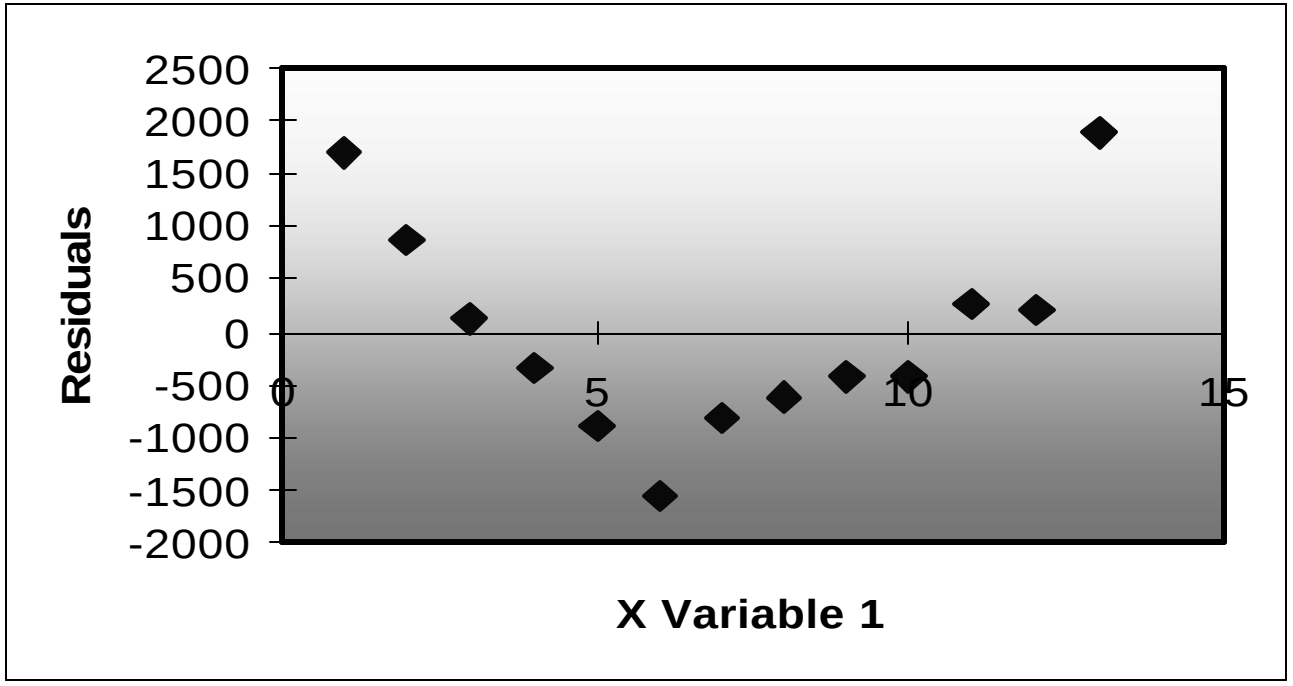
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3445.41	619.37	5.56	0.00	2082.18	4808.64
X Variable 1	270.02	78.03	3.46	0.01	98.27	441.77

Assumptions of Regression

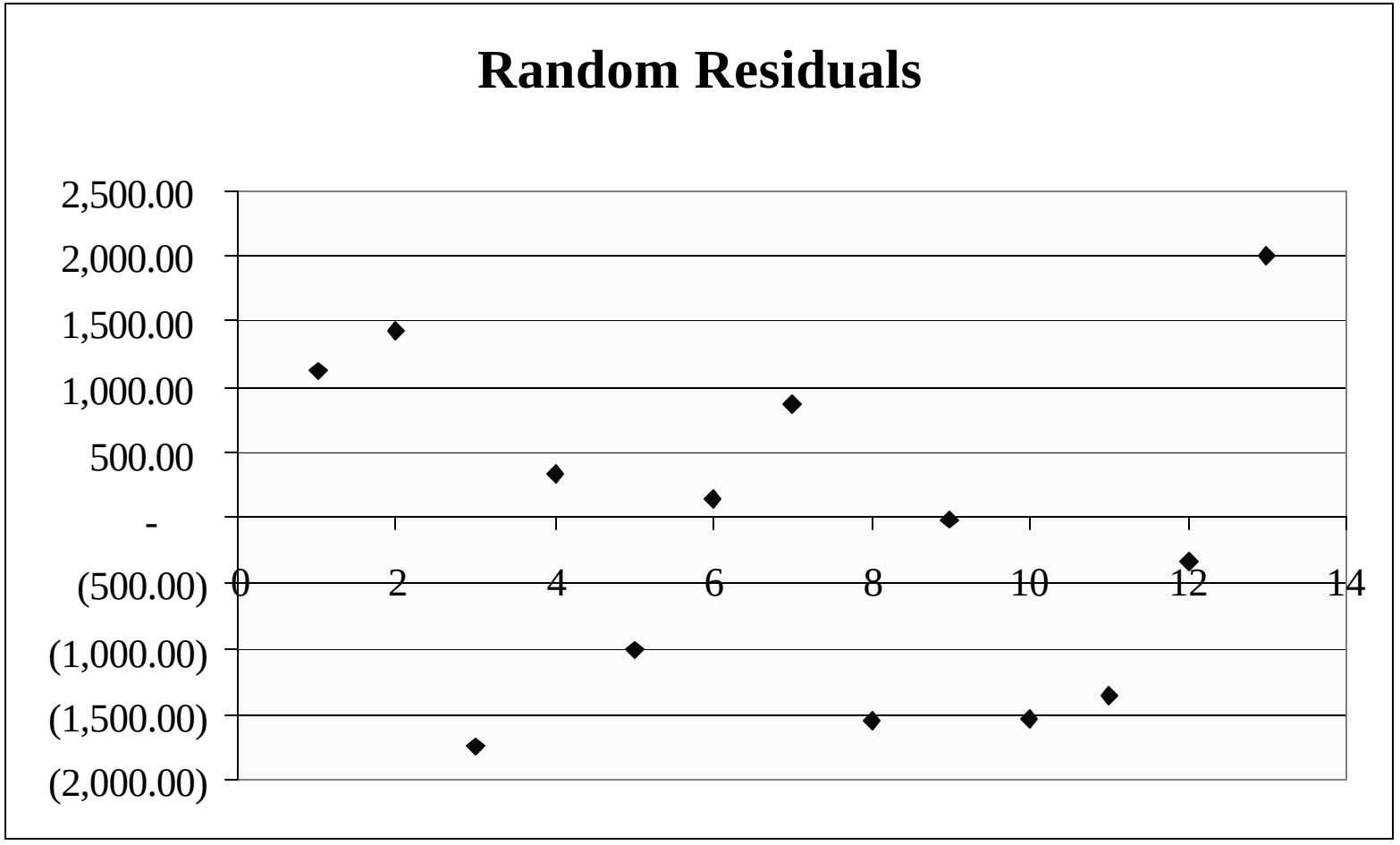
- Errors independent of value of X
- Errors independent of value of Y
- Errors independent of prior errors
- Errors are from normal distribution
- Linearity
- We can test for validity of assumptions

Diagnostics: Residual Plot

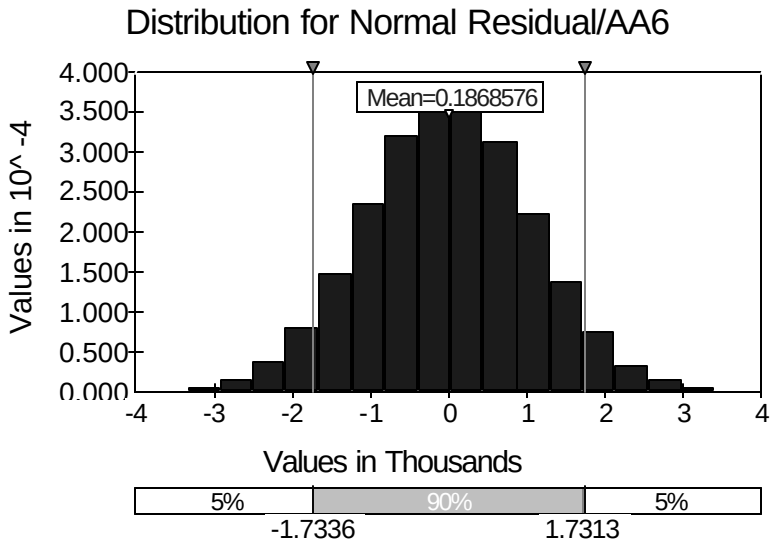
- Points should scatter randomly around zero
- If not, a straight line probably is not be appropriate



Random Residuals

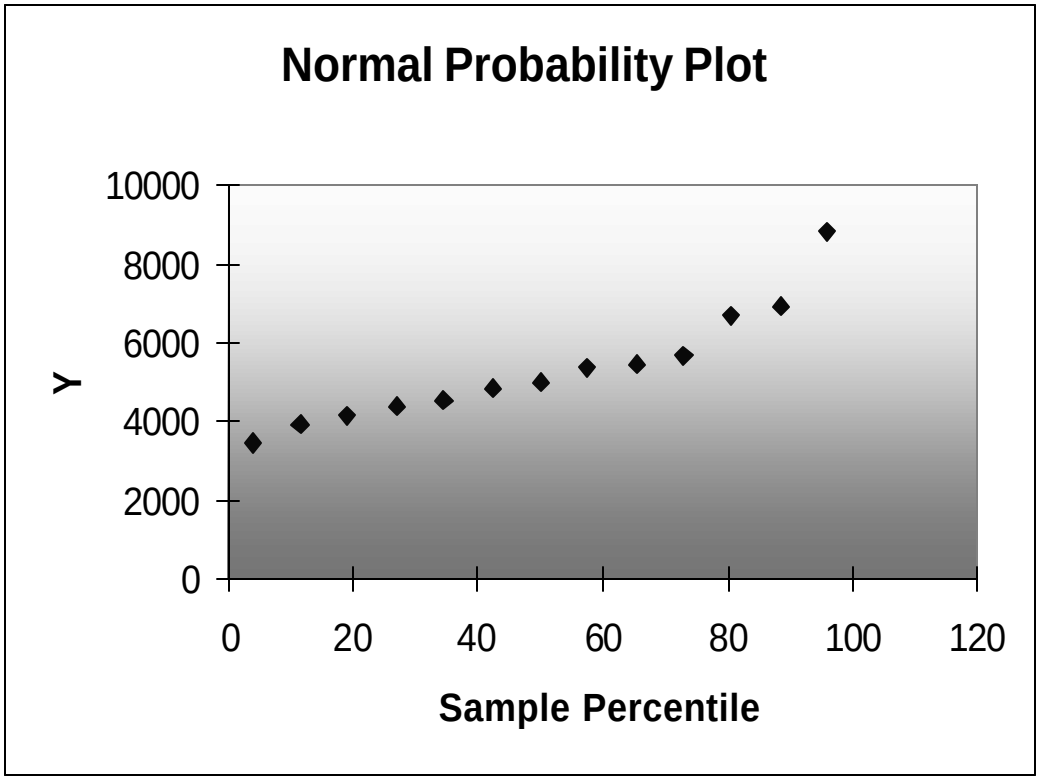


Normal Residual



Other Diagnostics: Normal QQ Plot

- Plot should be a straight line
- Otherwise residuals not from normal distribution



Test for autocorrelated errors

- Autocorrelation often present in time series data
- Durban – Watson statistic:
- If residuals uncorrelated, this is near 2

$$D = \frac{\sum_{i=2}^{N} (e_t - e_{t-1})^2}{\sum_{t-1}^{N} e_t^2}$$

Durban Watson Statistic

- Indicates autocorrelation present

(1)	(2)	(3)	(4)
Residual	Lag Residual	(e(t)-e(t-1))^2	e(t)^2
1,251.8			1,566,956.8
371.6	1,251.8	774,732.8	138,080.9
(425.6)	371.6	635,511.3	181,133.0
(527.5)	(425.6)	10,380.8	278,238.8
(824.2)	(527.5)	88,060.0	679,359.2
(1,087.3)	(824.2)	69,210.5	1,182,246.0
(140.7)	(1,087.3)	895,981.3	19,810.1
77.4	(140.7)	47,600.9	5,995.1
492.2	77.4	172,030.1	242,253.8
406.3	492.2	7,382.0	165,059.0
624.2	406.3	47,490.8	389,623.5
(199.6)	624.2	678,714.5	39,857.3
(18.5)	(199.6)	32,830.5	340.4
Sum		3,459,925.3	4,888,953.9
		DW =	0.707702582

Non-Linear Relationships

- The model fit was of the form:
  - Severity = a + b*Year
- A more common trend model is:
  - $Severity_{Year} = Severity_{Year0} * (1+t)^{(Year-Year0)}$
  - T is the trend rate
  - This is an exponential trend model
  - Cannot fit it with a line

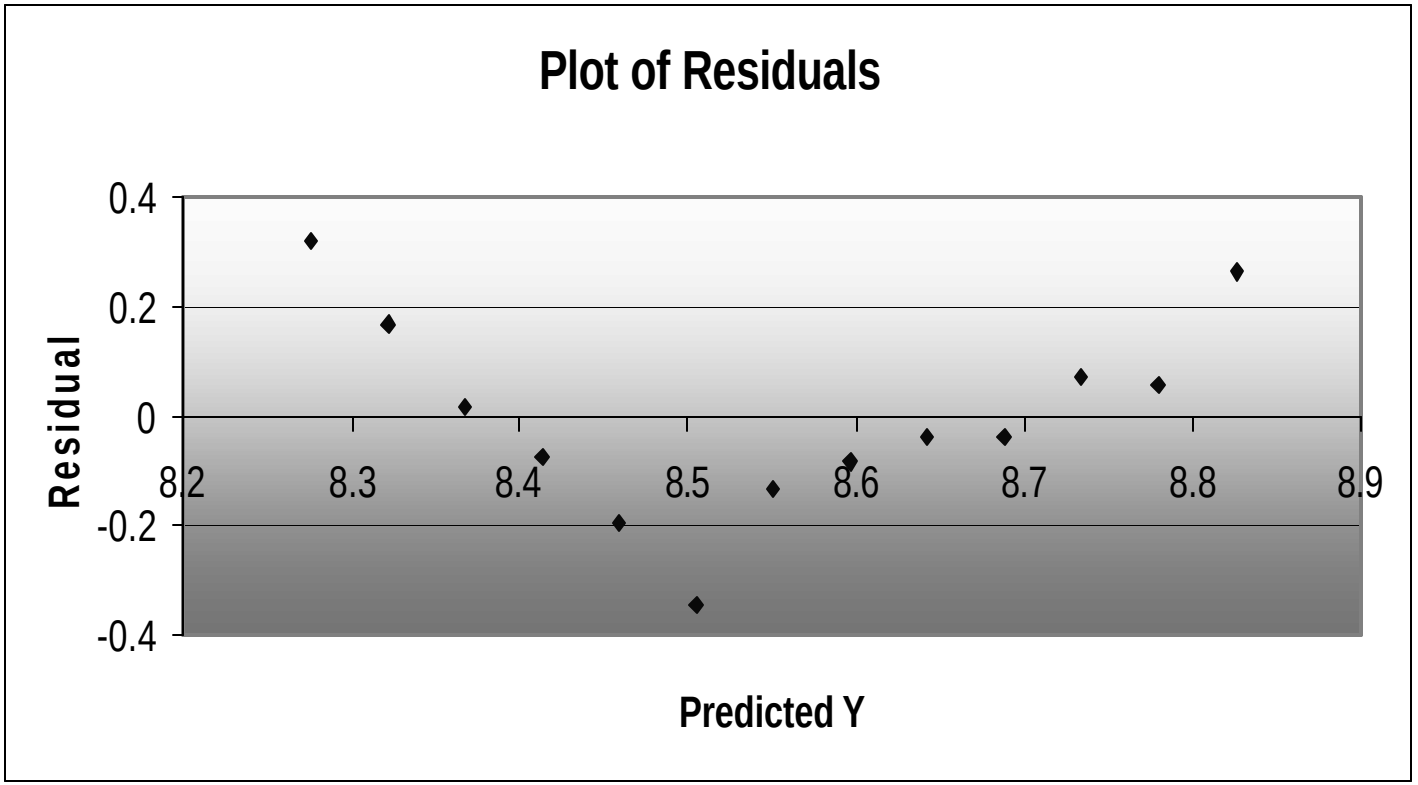
Transformation of Variables

- $Severity_{Year} = Severity_{Year0} * (1+t)^{(Year-Year0)}$

1. Log both sides
2. $\ln(Sev_{Year}) = \ln(Sev_{Year0}) + (Year-Year0) * \ln(1+t)$
3. Y = a + x * b
4. A line can be fit to transformed variables where dependent variable is log(Y)

Exponential Trend – Cont.

- R^2 declines and residuals indicate poor fit



A More Complex Model

- Use more than one variable in model (Econometric Model)
- In this case we use a medical cost index, the consumer price index and employment data (number employed, unemployment rate, change in number employed, change in UEP rate to predict workers compensation severity

Data							
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Year	WC Severity	Health Ins Index	CPI	Employment	changeEm	UEP Rate	Cng UEP
1991	5,410	11.7	136.2	117,718	-0.9%	6.8	1.2
1992	4,868	12.7	140.3	118,492	0.7%	7.5	1.1
1993	4,393	13.6	144.5	120,259	1.5%	6.9	0.9

Multivariable Regression

$$b = (X^T X)^{-1} X^T Y$$

$$\hat{Y} = X^T b$$

$$Variance = s^2 \left(X^T X \right)^{-1}$$

One Approach: Regression With All Variables

- Many variables not significant
- Over-parameterization issue
- How get best fitting most parsimonious model?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98808656
R Square	0.97631505
Adjusted R Sq	0.952630099
Standard Error	317.0246366
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	24857347.33	4142891.2	41.220903	0.000128191
Residual	6	603027.7213	100504.62		
Total	12	25460375.05			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	(46,924.96)	25,285.89	(1.86)	0.113	(108,797.36)	14,947.43
Ins Index	945.52	297.55	3.18	0.019	217.43	1,673.61
CPI	(523.23)	93.98	(5.57)	0.001	(753.19)	(293.27)
Employment	0.92	0.29	3.17	0.019	0.21	1.62
PchangeEmp	(42,369.42)	39,355.61	(1.08)	0.323	(138,669.21)	53,930.37
UEP Rate	743.58	760.09	0.98	0.366	(1,116.30)	2,603.46
Cng UEP	(633.27)	3,870.51	(0.16)	0.875	(10,104.09)	8,837.54

Multiple Regression Statistics

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.98808656
R Square	0.97631505
Adjusted R Sqı	0.952630099
Standard Error	317.0246366
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	24857347.33	4142891.2	41.220903	0.000128191
Residual	6	603027.7213	100504.62		
Total	12	25460375.05			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	(46,924.96)	25,285.89	(1.86)	0.113	(108,797.36)	14,947.43
Ins Index	945.52	297.55	3.18	0.019	217.43	1,673.61
CPI	(523.23)	93.98	(5.57)	0.001	(753.19)	(293.27)
Employment	0.92	0.29	3.17	0.019	0.21	1.62
PchangeEmp	(42,369.42)	39,355.61	(1.08)	0.323	(138,669.21)	53,930.37
UEP Rate	743.58	760.09	0.98	0.366	(1,116.30)	2,603.46
Cng UEP	(633.27)	3,870.51	(0.16)	0.875	(10,104.09)	8,837.54

Degrees of Freedom

- Related to number of observations
- One rule of thumb: subtract the number of parameters estimated from the number of observations
- The greater the number of parameters estimated, the lower the number of degrees of freedom

Degrees of Freedom

- “Degrees of freedom for a particular sum of squares is the smallest number of terms we need to know in order to find the remaining terms and thereby compute the sum”
  - Iverson and Norpoth, *Analysis of Variance*
- We want to keep the df as large as possible to avoid overfitting
- This concept becomes particularly important with complex data mining models

Stepwise Regression

- Partial correlation
  - Correlation of dependent variable with predictor after all other variables are in model
- F – contribution
  - Amount of change in F-statistic when variable is added to model

Stepwise regression-kinds

- Forward stepwise
  - Start with best one variable regression and add
- Backward stepwise
  - Start with full regression and delete variables
- Exhaustive

Stepwise regression for Severity Data

Variables Entered/Removed[a]

		Variables Entered	Variables Removed	Method
Model	1	*Health Ins Index*	.	*Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).*
	2	*Change in UEP*	.	*Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).*

a. Dependent Variable: Severity

Stepwise Regression-Excluded Variables

Excluded Variables[c]

Model			Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
Model	1	CPI	-.787[a]	-1.907	.086	-.517	.113
		Employment	-.417[a]	-1.353	.206	-.393	.233
		Pct Change in Employmeny	-.318[a]	-2.472	.033	-.616	.984
		Unemployment Rate	.243[a]	1.572	.147	.445	.882
		Change in UEP	.328[a]	2.497	.032	.620	.935
		Lag Employment	-.364[a]	-1.156	.274	-.343	.233
		Lag UEP	.123[a]	.761	.464	.234	.941
	2	CPI	-.555[b]	-1.484	.172	-.443	.103
		Employment	-.235[b]	-.835	.425	-.268	.210
		Pct Change in Employmeny	-.150[b]	-.408	.693	-.135	.129
		Unemployment Rate	.114[b]	.732	.483	.237	.703
		Change in UEP					
		Lag Employment	-.203[b]	-.726	.486	-.235	.217
		Lag UEP	.040[b]	.282	.784	.094	.876

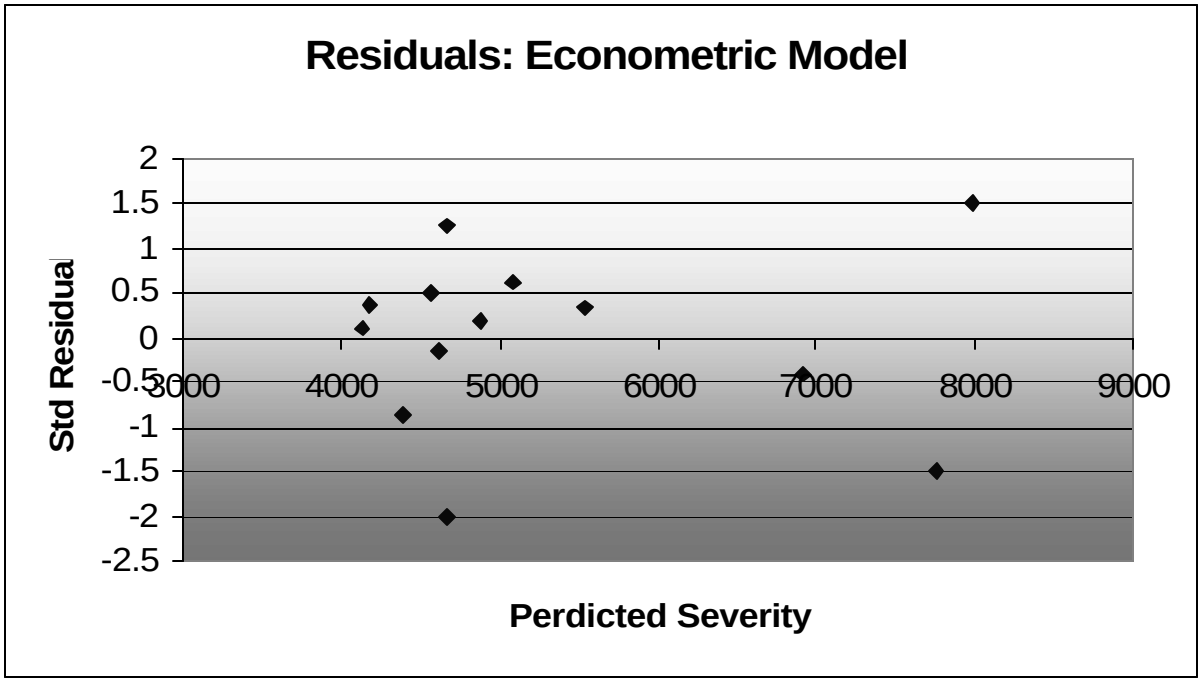
a. Predictors in the Model: (Constant), Health Ins Index

b. Predictors in the Model: (Constant), Health Ins Index, Change in UEP

c. Dependent Variable: Severity

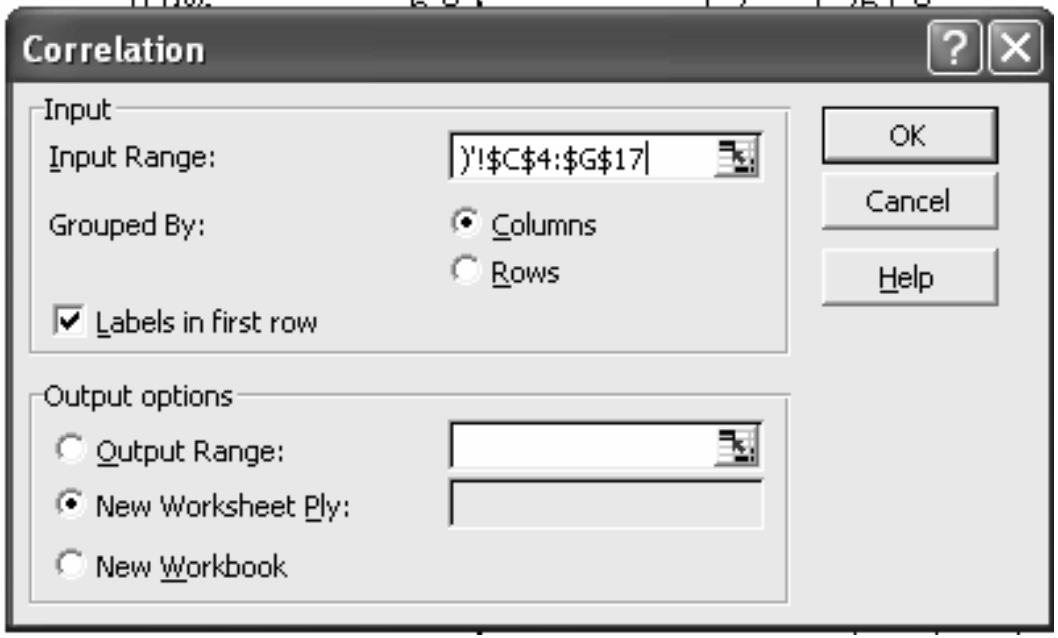
Econometric Model Assessment

- Standardized residuals more evenly spread around the zero line
- R^2 is .91 vs .52 of simple trend regression



Correlation of Predictor Variables: Multicollinearity

Ins Index	CPI	Employment	PchangeEmp	UEP Rate	Cng UEP	Residual	Resi
11.7	136.2	117,718	0.0%	6.9	1.2	1,251.9	
12.7	140.3	118,492					
13.6	144.5	120,259					
13.8	148.3	123,060					
14.3	152.4	124,900					
14.5	156.9	126,708					
15.1	160.6	129,558					
15.7	163.0	131,463					
16.1	166.6	133,488					
17.3	172.2	136,891					
18.9	177.1	136,933					
20.7	179.9	136,485					
23.6	184.0	137,736					



Multicollinearity

- Predictor variables are assumed uncorrelated
- Assess with correlation matrix

	Ins Index	*CPI*	*Employment*	*PchangeEmp*	*UEP Rate*	*Cng UEP*
Ins Index	1.000					
CPI	0.942	1.000				
Employment	0.876	0.984	1.000			
PchangeEmp	(0.125)	0.016	0.092	1.000		
UEP Rate	(0.344)	(0.622)	(0.742)	(0.419)	1.000	
Cng UEP	0.254	0.143	0.077	(0.926)	0.321	1.000

Remedies for Multicollinearity

- Drop one or more of the highly correlated variables
- Use Factor analysis or Principle components to produce a new variable which is a weighted average of the correlated variables
- Use stepwise regression to select variables to include

Exponential Smoothing

- A weighted average with more weight given to more recent values

$$\hat{Y}_t = a\hat{Y}_{t-1} + (1-a)Y_t$$

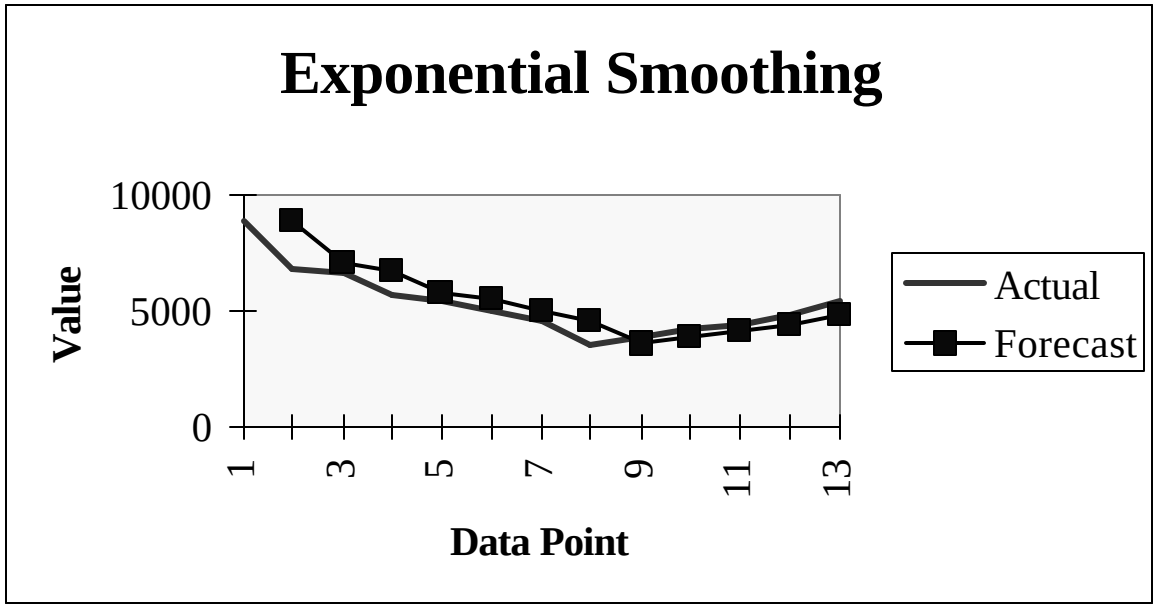
- Linear Exponential Smoothing: model level and trend

a usually between .05 and .3

$$m_t = dY_t + (1-d)(m_{t-1} + r_{t-1})$$

$$r_t = a(m_t - m_{t-1}) + (1-a)r_{t-1}$$

Exponential Smoothing Fit



Tail Development Factors: Another Regression Application (in WC Claims Pred Seminar.xls)

- Typically involve non-linear functions:
  - Inverse Power Curve:

$$LDF_t = 1 + \frac{k}{(t+c)^a}$$

  - Hoerel Curve:
  - Probability distribution such as Gamma, Lognormal

$$LDF_t = Kt^{a-1}\exp(-bt)$$

$$LDF_t = \frac{F(t+1)}{F(t)}$$

Example: Inverse Power Curve

- Can use transformation of variables to fit simplified model: $LDF=1+k/t^a$
  - $\ln(LDF-1) = a+b*\ln(1/t)$
- Use nonlinear regression to solve for a and c
  - Uses numerical algorithms, such as gradient descent to solve for parameters.
  - Most statistics packages let you do this

Nonlinear Regression: Grid Search Method

- Try out a number of different values for parameters and pick the ones which minimize a goodness of fit statistic
- You can use the Data Table capability of Excel to do this
  - Use regression functions linest and intercept to get k and a
  - Try out different values for *c* until you find the best one

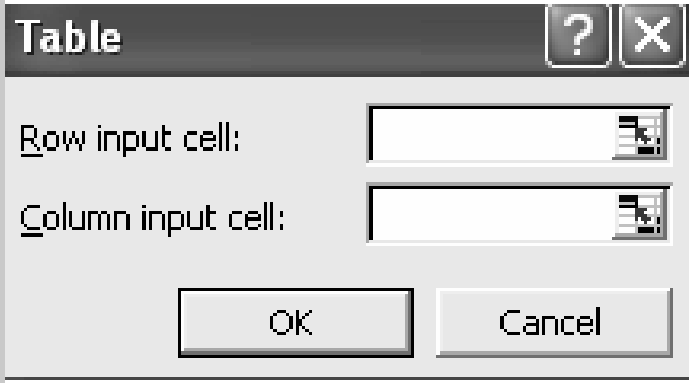
Fitting non-linear function

		c= -5			
Age	WC LDF	ln(ldf-1)	ln(1/(t+c)	Fitted	Error^2
12	1.502	-0.68916	-1.94591	1.511531	9.08E-05
24	1.148	-1.91054	-2.94444	1.117107	0.000954
36	1.063	-2.76462	-3.43399	1.056842	3.79E-05
48	1.028	-3.57555	-3.7612	1.035063	4.99E-05
60	1.017	-4.07454	-4.00733	1.024379	5.45E-05
72	1.013	-4.34281	-4.20469	1.018217	2.72E-05
84	1.019	-3.96332	-4.36945	1.014283	2.23E-05
96	1.014	-4.2687	-4.51086	1.011592	5.8E-06
108	1.011	-4.50986	-4.63473	1.009654	1.81E-06
					0.001245

LDFs from www.njcrib.org

Coefficient	1.476494	Linest
Constant	2.202778	Intercept

Using Data Tables in Excel



c	0.001245 ← error
-7	0.00462
-5	0.00124
-3	0.00141
-2	0.00219
-1	0.00323
0	0.00446
1	0.00580
2	0.00723
5	0.01168
10	0.01887

Use Model to Compute the Tail

(Using Prediction through 50 years of Development)

(1)	(2)	(3)	(4)
		EXP(a+b*(2))	
Age	ln(1/(t+c)	Prediced Tail Factors	Cumulative Tail
120	-4.74493	1.0082	1.1054
132	-4.84419	1.0071	1.0964
144	-4.93447	1.0062	1.0887
156	-5.01728	1.0055	1.0820
168	-5.09375	1.0049	1.0761
180	-5.16479	1.0044	1.0709
192	-5.23111	1.0040	1.0661
204	-5.2933	1.0037	1.0619
216	-5.35186	1.0033	1.0580
228	-5.40717	1.0031	1.0545
240	-5.45959	1.0029	1.0513
252	-5.50939	1.0027	1.0483
264	-5.55683	1.0025	1.0455
276	-5.60212	1.0023	1.0429

Fitting Non-linear functions

- Another approach is to use a numerical method
  - Newton-Raphson (one dimension)
    - $x_{n+1} = x_n - f'(x_n)/f''(x_n)$
    - $f(x_n)$ is typically a function being maximized or minimized, such as squared errors
    - x's are parameters being estimated
  - A multivariate version of Newton-Raphson or other algorithm is available to solve non-linear problems in most statistical software
  - In Excel the Solver add-in is used to do this

Claim Count Triangle Model

- Chain ladder is common approach

Workers Compensation Claim Counts

YEAR	AT 12 MONTHS	AT 24 MONTHS	AT 36 MONTHS	AT 48 MONTHS	AT 60 MONTHS	Reported as of 12/31/03	Loss Development Factor	Ultimate Claims
1998	112	134	136	136	136	136	1.000	136.0
1999	78	106	110	110	110	110	1.000	110.0
2000	68	101	101	101		101	1.000	101.0
2001	101	123	123			123	1.000	123.0
2002	113	124				124	1.013	125.6
2003	114					114	1.266	144.4

YEAR	12-24	24-36	36-48
1998	1.196	1.015	1.000
1999	1.359	1.038	1.000
2000	1.485	1.000	1.000
2001	1.218	1.000	
2002	1.097		
Average	1.271	1.013	1.000
Wt Avg	1.246	1.013	1.000
Selected	1.250	1.013	1.000
Age to Ultimate	1.266	1.013	1.000

Claim Count Development

- Another approach: additive model

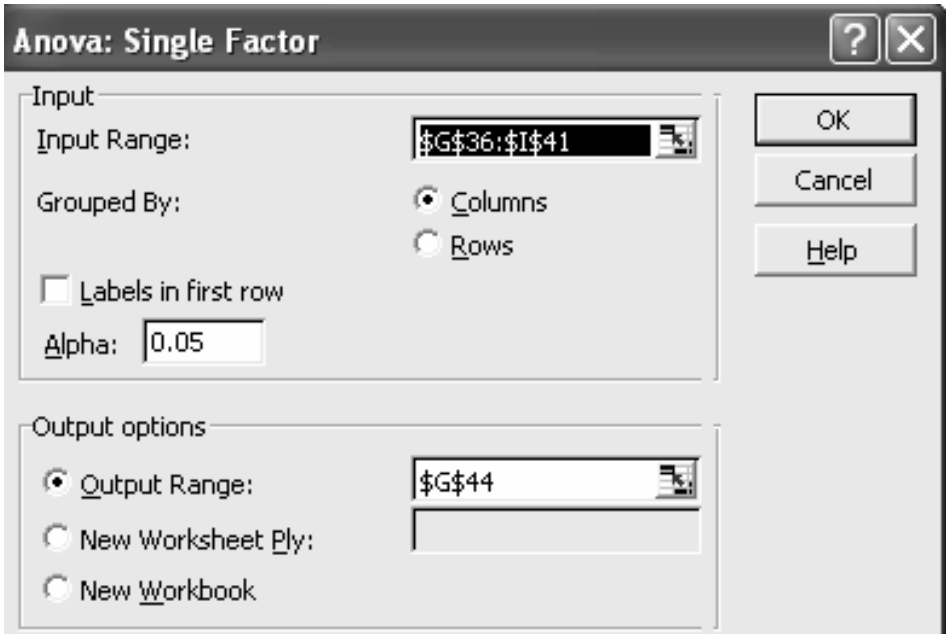
$$Y_{i,j} = m_j + e$$

$Y_{i,j}$ = incremental claims

- This model is the same as a one factor ANOVA

ANOVA Model for Development

----------Actual Claims----------		
12	24	36
168	25	1
117	33	0
102	42	0
185	50	3
170	0	6
171	16	0



ANOVA: Two Groups

- With only two groups we test the significance of the difference between their means
- Many years ago in a statistics course we learned how to use a t-test for this

$$s = \sqrt{s^2/n_1 + s^2/n_2}$$

$$t = \frac{\ddot{y}_1 - \ddot{y}_2}{s}$$

$$F_{1,n_1+n_2-2} = \frac{n_2(\bar{y}_2 - \bar{y})^2 + n_1(\bar{y}_1 - \bar{y})^2}{s^2}$$

ANOVA: More than 2 Groups

$$F_{2,n_1+n_2+n_3-3} = \frac{n_3(\bar{y}_2 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + n_1(\bar{y}_1 - \bar{y})^2}{s^2}$$

Correlation Measure: Eta

$$h = \sqrt{\frac{SS_{between}}{SS_{total}}}$$

ANOVA Model for Development

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	6	913	152.17	1150.97
Column 2	6	166	27.67	328.27
Column 3	6	10	1.67	5.87

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	77653	2	38826.50	78.43	0.00	3.68
Within Groups	7425.5	15	495.03			
Total	85078.5	17				

Regression With Dummy Variables

- Let Devage24=1 if development age = 24 months, 0 otherwise
- Let Devage36=1 if development age = 36 months, 0 otherwise
- Need one less dummy variable than number of ages

Regression with Dummy Variables: Design Matrix

YEAR	Age	Cumulative	Claims	Devage24	Devage36
1998	1	168	168	0	0
1999	1	117	117	0	0
2000	1	102	102	0	0
2001	1	185	185	0	0
2002	1	170	170	0	0
2003	1	171	171	0	0
1997	2	99	25	1	0
1998	2	201	33	1	0
1999	2	159	42	1	0
2000	2	152	50	1	0
2001	2	185	0	1	0
2002	2	186	16	1	0
1995	3	140	1	0	1
1996	3	121	0	0	1
1997	3	99	0	0	1
1998	3	204	3	0	1
1999	3	165	6	0	1
2000	3	152	0	0	1

Equivalent Model to ANOVA

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.955365
R Square	0.912722
Adjusted R Square	0.901085
Standard Error	22.24934
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	77653	38826.5	78.43209	1.13971E-08
Residual	15	7425.5	495.0333		
Total	17	85078.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	152.1667	9.083257	16.75243	4.04E-11	132.806151	171.5272
X Variable 1	-124.5	12.84567	-9.69199	7.53E-08	-151.8799038	-97.1201
X Variable 2	-150.5	12.84567	-11.716	5.99E-09	-177.8799038	-123.12
Age 2 estimate	27.7					
Age 3 estimate	1.7					

Apply Logarithmic Transformation

- It is reasonable to believe that variance is proportional to expected value
- Claims can only have positive values
- If we log the claim values, can’t get a negative
- Regress log(Claims+.001) on dummy variables or do ANOVA on logged data

Log Regression

YEAR	Age	Cumulative	Claims	ln(Claims)	DevAge24	Devage36
1998	1	168	168	5.12397	0	0
1999	1	117	117	4.762182	0	0
2000	1	102	102	4.624983	0	0
2001	1	185	185	5.220361	0	0
2002	1	170	170	5.135804	0	0
2003	1	171	171	5.141669	0	0
1997	2	99	25	3.218916	1	0
1998	2	201	33	3.496538	1	0
1999	2	159	42	3.737693	1	0
2000	2	152	50	3.912043	1	0
2001	2	185	0	-6.90776	1	0
2002	2	186	16	2.772651	1	0
1995	3	140	1	0.001	0	1
1996	3	121	0	-6.90776	0	1
1997	3	99	0	-6.90776	0	1
1998	3	204	3	1.098946	0	1
1999	3	165	6	1.791926	0	1
2000	3	152	0	-6.90776	0	1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.714499
R Square	0.510508
Adjusted R	0.445243
Standard E	3.509039
Observatio	18

ANOVA

	df	*SS*	*MS*
Regressior	2	192.6306	96.31532
Residual	15	184.7003	12.31335
Total	17	377.3309	

	Coefficients	*tandard Err*	*t Stat*
Intercept	5.001495	1.432559	3.491301
X Variable	-3.29648	2.025945	-1.62713
X Variable	-7.97339	2.025945	-3.93564

Poisson Regression

- Log Regression assumption: errors on log scale are from normal distribution.
- But these are claims – Poisson assumption might be reasonable
- Poisson and Normal from more general class of distributions: exponential family of distributions

“Natural” Form of the Exponential Family

$$f(y_i;q_i,f)=\exp\left[\frac{\{q_i\cdot y_i-b(q_i)\}}{a(f)}+c(y_i,f)\right]$$

Specific Members of the Exponential Family

- Normal (Gaussian)
- Poisson
- Negative Binomial
- Gamma
- Inverse Gaussian

Some Other Members of the Exponential Family

- Natural Form
  - Binomial
  - Compound Poisson/Gamma (Tweedie)
- General Form [use ln(y) instead of y]
  - Lognormal
  - Single Parameter Pareto

Poisson Distribution

- Poisson distribution:

$$\Pr ob(Y = y) = \frac{m^y}{y!} e^{-m}$$

$$\Pr ob(Y = y) = \exp(\ln(\frac{m^y}{y!}) - m)$$

$$\Pr ob(Y = y) = \underbrace{\exp(y \ln m)}_{yq} - \underbrace{\ln(y!)}_{c(y)} - \underbrace{m}_{b(q)}$$

Poisson Distribution

- Poisson distribution: $\Pr ob(Y = y) = \frac{m^y}{y!} e^{-m}$

- Natural Form:

$$\text{Prob}(Y = y) = \exp\left[\frac{\{\ln(m) \cdot y - m\}}{f} - y \cdot \frac{\ln(f)}{f} - \ln((y/f)! \right.$$

- "Over-dispersed" Poisson allows $f \neq 1$.
  - Variance/Mean ratio = f

Linear Model vs GLM

- Regression:

$$Y_i = m_i + e$$

$$m_i = X'\mathrm{B}$$

$$e \sim N(0, s^2)$$

- GLM:

$$Y = h(m_i) + e$$

$$h(m) = X'\mathrm{B}$$

$$e \sim \text{exponential family}$$

h is a link function

The Link Function

- Like transformation of variables in linear regression
  - $Y=AX^B$ is transformed into a linear model
    - log(Y) = log(A) + B*log(X)
  - This is similar to having a log link function:
    - h(Y) = log(Y)
    - denote h(Y) as n
    - n = a+bx

Other Link Functions

- Identity
  - h(Y)=Y
- Inverse
  - h(Y) = 1/Y
- Logistic
  - h(Y)=log(y/(1-y))
- Probit
  - h(Y) =

$$\Phi(Y), \Phi \text{ is Normal Distribution}$$

The Other Parameters: Poisson Example

exponential family

$$f(y_i;\ q_i,\ f) = \exp\left[\frac{\{q_i \cdot y_i - b(q_i)\}}{a(f)} + c(y_i, f)\right]$$

Poisson:

$$\text{Prob}(Y = y) \quad = \quad \exp\left[\frac{\{\ln(m)\cdot y - m\}}{f} - y\cdot\frac{\ln(f)}{f} - \ln((y/f)!)\right]$$

$$E(y) = b(q) = \exp(q) = m$$

so $q = \ln(m)$ ← Link function

$$\text{var}(Y) = b''(q)a(f)$$

$b''(q)$ is variance function and equals m f for poisson

$$a(f) = \frac{f}{w}$$

with standard Poisson f and w are 1, but more about them late

LogLikhood for Poisson

$$f(y) = \frac{m^y}{y!}\exp(-m)$$

$$L(y) = \prod_{i=1}^{N} \frac{m^{yi}}{y_i!}\exp(-m)$$

$$\log(L(y)) = \sum_{i=1}^{N} y_i \ln(m) - m$$

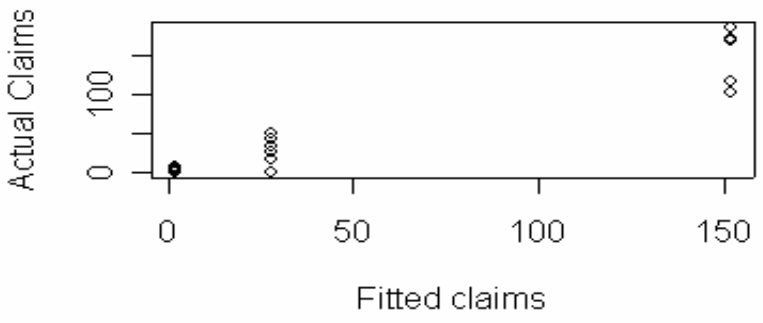
with regression: $l\,\mathrm{og}(L(y)) = \sum_{i=1}^{N} y_i \ln(a + b_1 x \ldots b_n x) - (a + b_1 x \ldots b_n x)$

Estimating Parameters

- As with nonlinear regression, there usually is not a closed form solution for GLMs
- A numerical method used to solve
- For some models this could be programmed in Excel – but statistical software is the usual choice
- If you can't spend money on the software, download R for free

GLM fit for Poisson Regression

```
>devage<-as.factor((AGE)
>claims.glm<-glm(Claims~devage, family=poisson)
>summary(claims.glm)
Call:
glm(formula = Claims ~ devage, family = poisson)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.250   -1.732   -0.500    0.507   10.626
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.73540    0.02825 167.622  < 2e-16 ***
devage2     -0.89595    0.05430 -16.500  < 2e-16 ***
devage3     -4.32994    0.29004 -14.929  < 2e-16 ***
devage4     -6.81484    1.00020  -6.813 9.53e-12 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 2838.65  on 36  degrees of freedom
Residual deviance:  708.72  on 33  degrees of freedom
AIC: 851.38
```



Deviance: Testing Fit

- The maximum liklihood achievable is a full model with the actual data, y_i, substituted for E(y)
- The liklihood for a given model uses the predicted value for the model in place of E(y) in the liklihood
- Twice the difference between these two quantities is known as the deviance
- For the Normal, this is just the errors
- It is used to assess the goodness of fit of GLM models – thus it functions like residuals for Normal models

A More General Model for Claim Development

$$Y_{i,j} = m_i + m_j + e$$

or Multiplicative model

$$Y_{i,j} = Bm_i m_j + e$$

m_i is accident year effect, m_j is development age effect

$Y_{i,j}$ = incremental claims

Design Matrix: Dev Age and Accident Year Model

YEAR	Incremental Claims	DevAge24	Devage36	Devage48	AY94	AY95	AY96	AY97
1993	27	0	0	0	0	0	0	0
1993	136	0	0	0	0	0	0	0
1993	1	0	0	0	0	0	0	0
1993	0	0	0	1	0	0	0	0
1994	24	0	0	0	1	0	0	0
1994	118	0	0	0	1	0	0	0
1994	1	0	0	0	1	0	0	0
1994	1	0	0	1	1	0	0	0
1995	116	0	0	0	0	1	0	0
1995	23	0	0	0	0	1	0	0
1995	1	0	0	0	0	1	0	0
1995	0	0	0	1	0	1	0	0
1996	99	0	0	0	0	0	1	0
1996	22	0	0	0	0	0	1	0
1996	0	0	0	0	0	0	1	0
1996	0	0	0	1	0	0	1	0
1997	74	0	0	0	0	0	0	1
1997	25	0	0	0	0	0	0	1

More General GLM development Model

Deviance Residuals:

Min	1Q	Median	3Q	Max
-10.5459	-1.4136	-0.4511	0.7035	10.2242

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.731366	0.079903	59.214	< 2e-16	***
devage2	-0.844529	0.055450	-15.230	< 2e-16	***
devage3	-4.227461	0.290609	-14.547	< 2e-16	***
devage4	-6.712368	1.000482	-6.709	1.96e-11	***
AY1994	-0.130053	0.114200	-1.139	0.254778	
AY1995	-0.158224	0.115066	-1.375	0.169110	
AY1996	-0.304076	0.119841	-2.537	0.011170	*
AY1997	-0.504747	0.127273	-3.966	7.31e-05	***
AY1998	0.218254	0.104878	2.081	0.037431	*
AY1999	0.006079	0.110263	0.055	0.956033	
AY2000	-0.075986	0.112589	-0.675	0.499742	
AY2001	0.131483	0.107294	1.225	0.220408	
AY2002	0.136874	0.107159	1.277	0.201496	
AY2003	0.410297	0.110600	3.710	0.000207	***

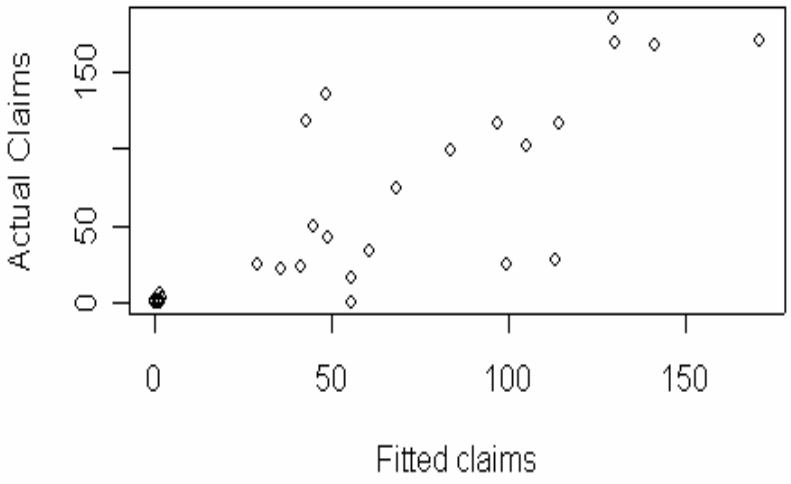
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for poisson family taken to be 1)

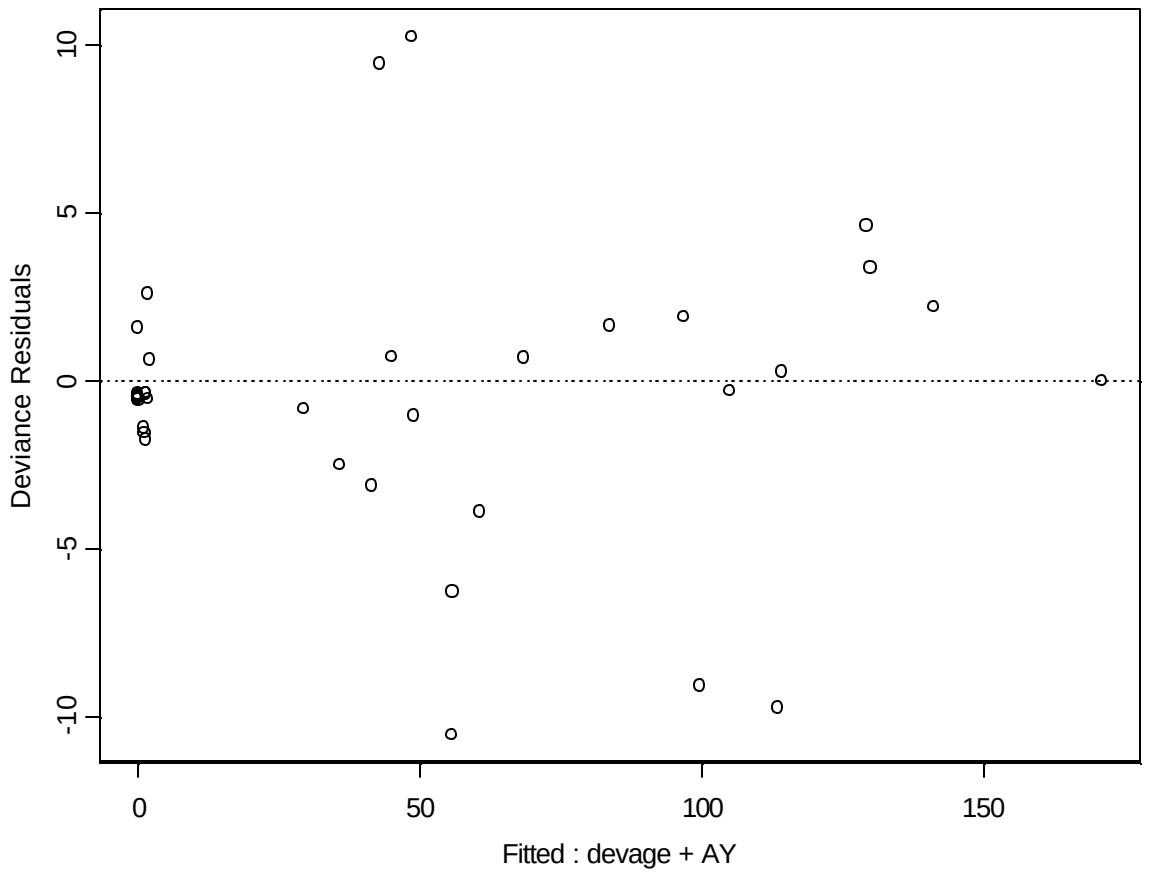
Null deviance: 2838.65 on 36 degrees of freedom

Residual deviance: 619.64 on 23 degrees of freedom

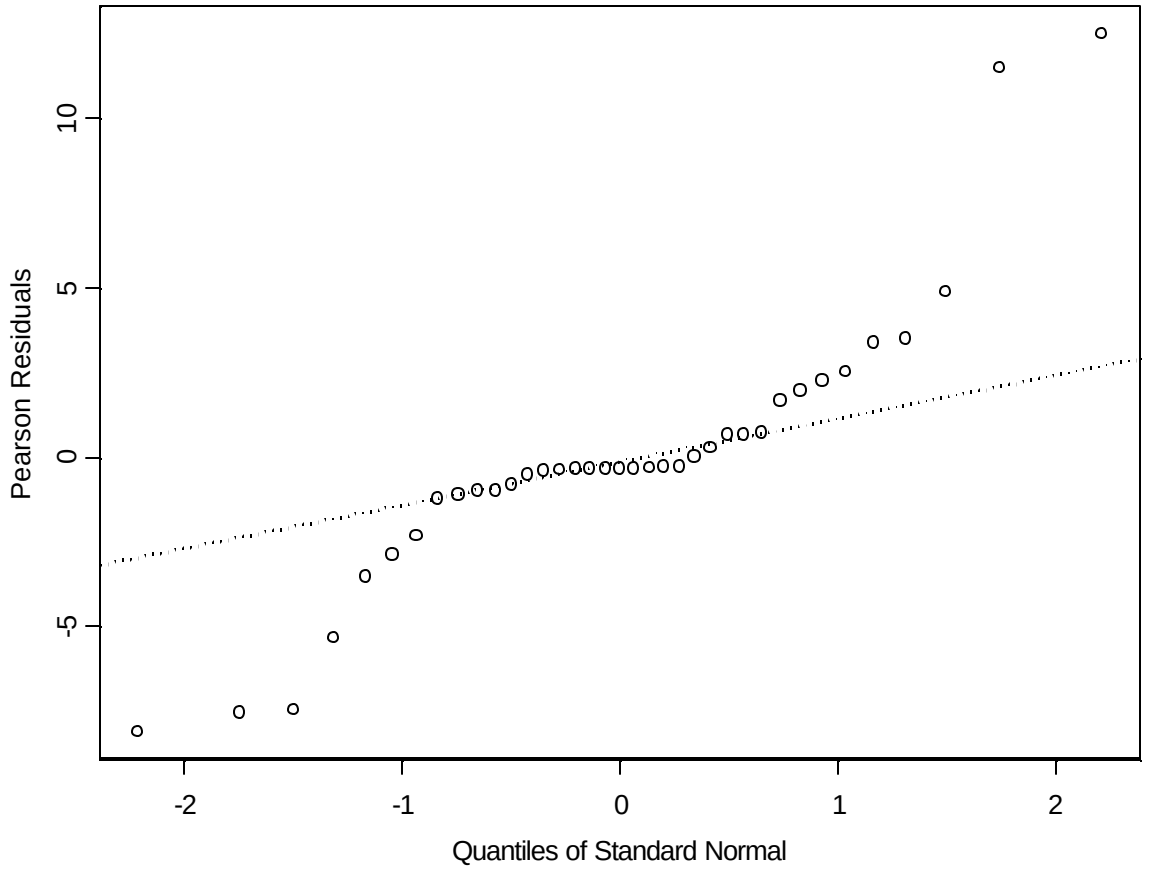
AIC: 782.3



Plot Deviance Residuals to Assess Fit



QQ Plots of Residuals



An Overdispersed Poisson?

- Variance of poisson should be equal to its mean
- If it is greater than that, then overdispersed poisson
- This uses the parameter
- It is estimated by evaluating how much the actual variance exceeds the mean

Weighted Regression

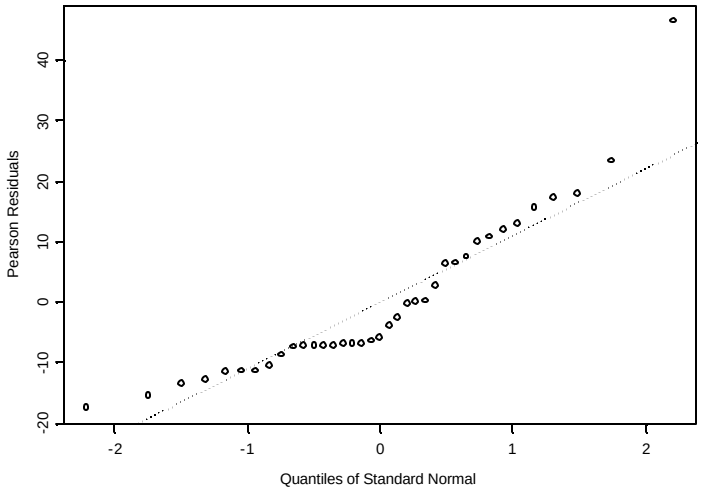
- There an additional consideration in the analysis: should the observations be weighted?
- The variability of a particular record will be proportional to exposures
- Thus, a natural weight is exposures

Weighted Regression

- Example:
  - Severities more credible if weighted by number of claims they are based on
  - Frequencies more credible if weighted by exposures
  - Weight inversely proportional to variance
  - Like a regression with # observations equal to number of claims (policyholders) in each cell
  - With GLM, specify appropriate weight variable in software

Weighted GLM of Claim Frequency Development

- Weighted by exposures
- Adjusted for overdispersion



Simulated Workers Compensation Data

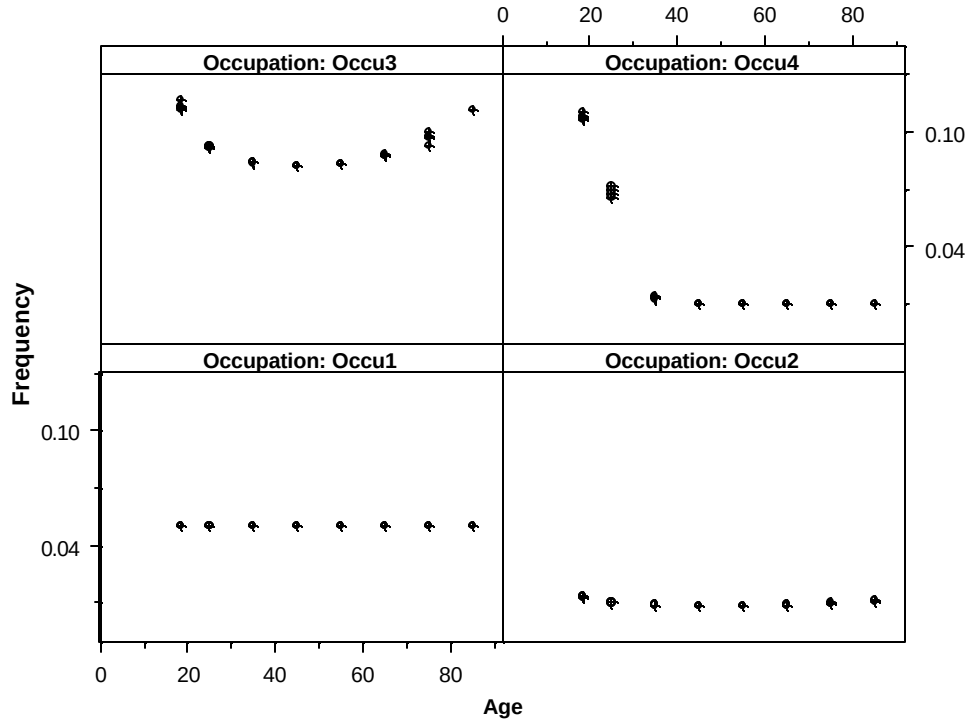
- Dependent variable
  - Frequency – claims per employee
- Predictor variables
  - Occupation
  - Age
  - Size of company
- Data was created to illustrate commonly encountered data complexities

Interactions

- Interactions occur when the relationship between a dependent and independent variable varies based on a third variable
- For instance suppose the relationship between age and the probability of a workers compensation injury varies by occupation

Example of Interaction

Frequency vs Age by Occupation



Interactions in Regression

- In regression, interactions are typically handled with interaction or product terms
- Interaction terms are like a combination of dummy variables and linear predictor variables
  - Let D1 by 1 if the employee is in occupation 1 and 0 otherwise
  - Let X be the employee's age
  - I1 = D1 * X
    - I1 is the interaction term

Regression with Interaction Terms

Frequency	Occupation	Dummy1	Dummy2	Dummy3	Age	Age*D1	Age*D1	Age*D1
5.7%	Occu1	1	0	0	18.5	18.5	0	0
4.4%	Occu1	1	0	0	25	25	0	0
4.4%	Occu1	1	0	0	35	35	0	0
3.9%	Occu1	1	0	0	45	45	0	0
4.0%	Occu1	1	0	0	55	55	0	0
1.9%	Occu1	1	0	0	65	65	0	0
13.3%	Occu1	1	0	0	75	75	0	0
0.0%	Occu1	1	0	0	85	85	0	0
3.0%	Occu2	0	1	0	18.5	0	18.5	0
1.7%	Occu2	0	1	0	25	0	25	0
2.5%	Occu2	0	1	0	35	0	35	0
0.8%	Occu2	0	1	0	45	0	45	0
1.3%	Occu2	0	1	0	55	0	55	0
0.0%	Occu2	0	1	0	65	0	65	0
0.0%	Occu2	0	1	0	75	0	75	0
13.8%	Occu3	0	0	1	18.5	0	0	18.5

Output of Regression with Interactions

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.64659
R Square	0.418079
Adjusted F	0.381709
Standard E	0.042525
Observatic	120

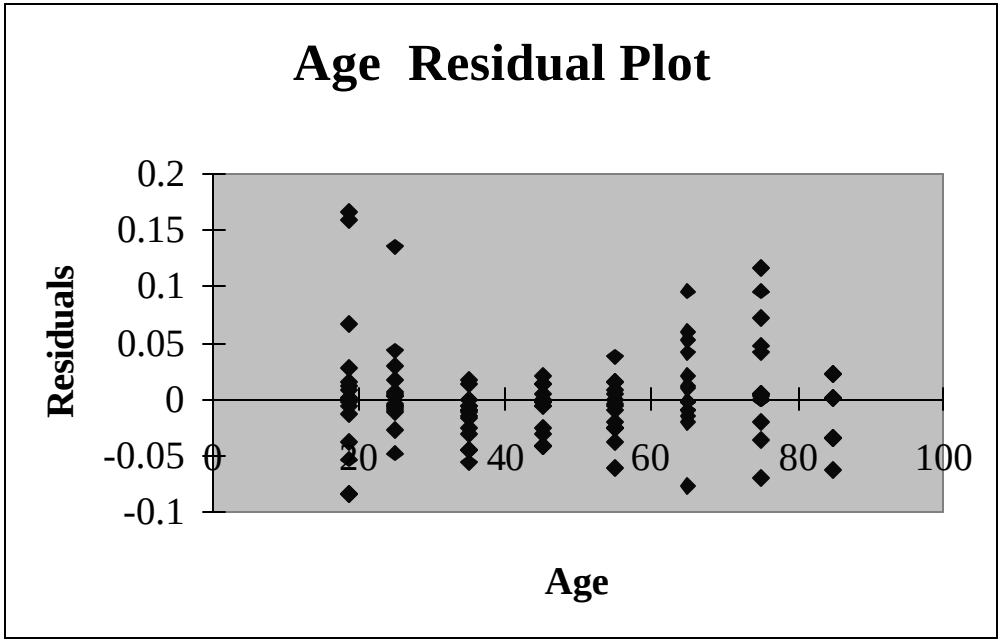
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	7	0.145514	0.020788	11.49515	6.23E-11
Residual	112	0.202539	0.001808		
Total	119	0.348053			

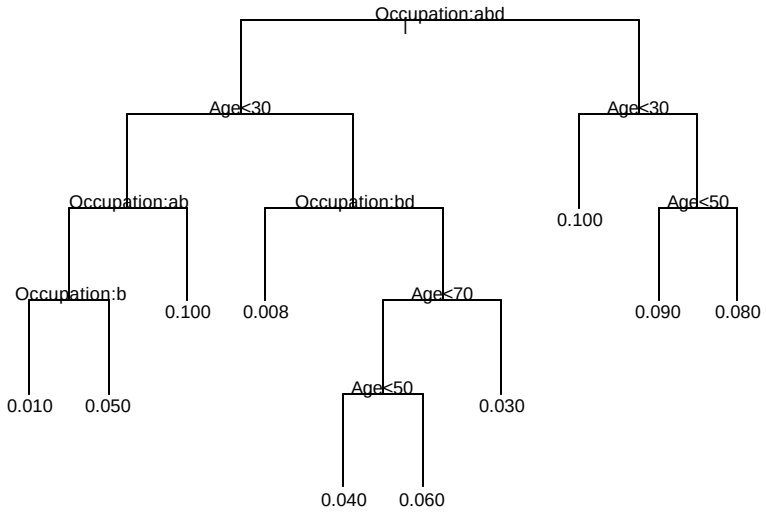
	Coefficients	*ndard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.113244	0.019364	5.848271	5E-08	0.074877	0.151611
Dummy1	-0.05146	0.027092	-1.89929	0.060098	-0.10514	0.002224
Dummy2	-0.09351	0.027384	-3.41473	0.00089	-0.14777	-0.03925
Dummy3	0.010024	0.027763	0.361047	0.718745	-0.04499	0.065033
Age	-0.00158	0.000369	-4.28458	3.89E-05	-0.00231	-0.00085
Age*D1	0.001254	0.000509	2.462199	0.015332	0.000245	0.002263
Age*D1	0.001324	0.000521	2.539389	0.012476	0.000291	0.002356
Age*D1	0.000879	0.000536	1.638323	0.104161	-0.00018	0.001941

Residual Plot of Interaction Regression

- Residuals indicate a pattern
- This is a result of non-linear relationships



Common Method for Modeling Interactions and Nonlinearities :Regression Trees



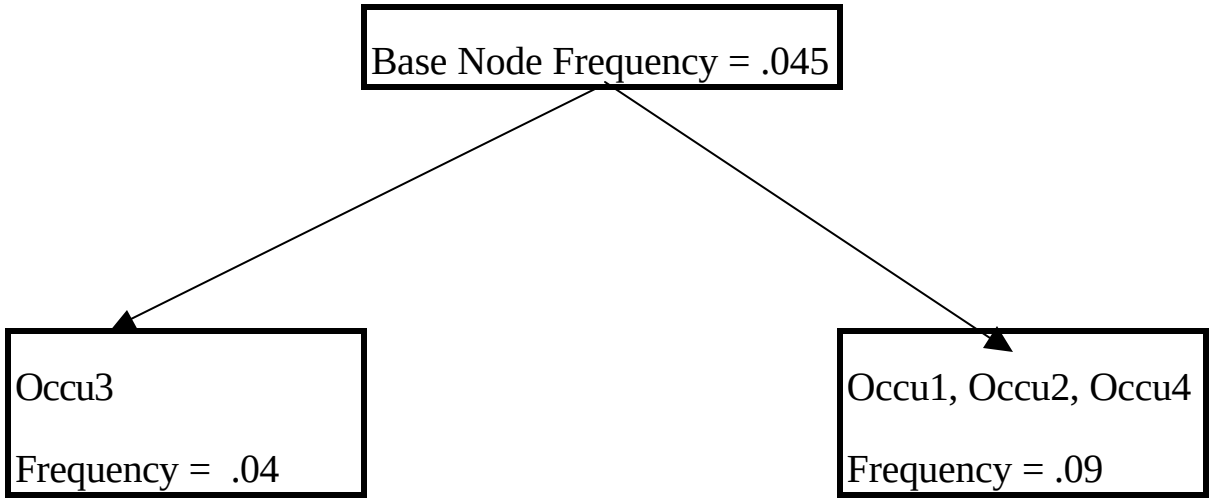
Regression Tree

- Based on sequential splitting of the data
- The first split creates the two groups with that produce the “best” split of the data
- R^2 or F is typically the goodness of fit test
- Regression Trees are essentially computationally intensive ANOVAs

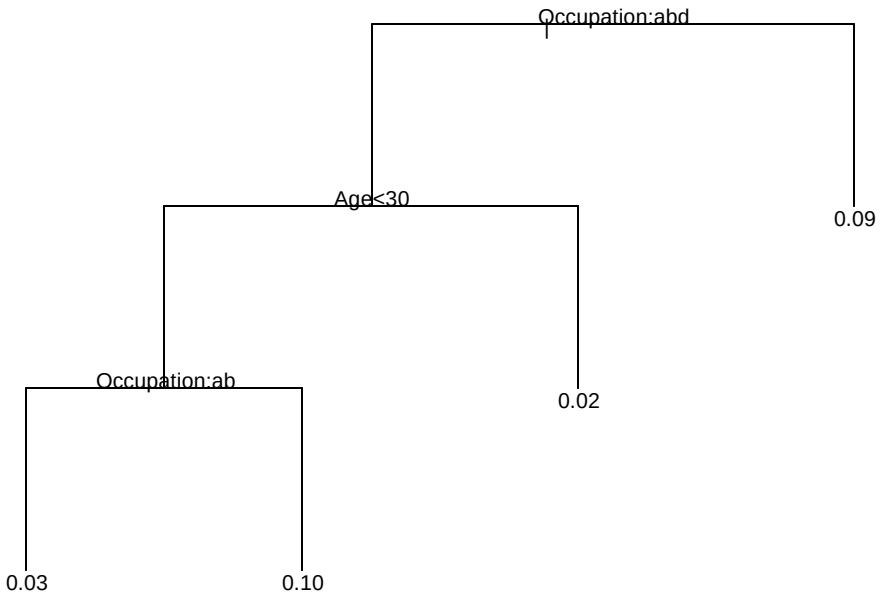
Regression Tree

Some Combinations of Occupation	R^2
Occu1 and Occu2	0.226295
Occu3 and Occu4	0.249184
Occu1	0.009998
Occu2	0.044531
Occu3	0.401094

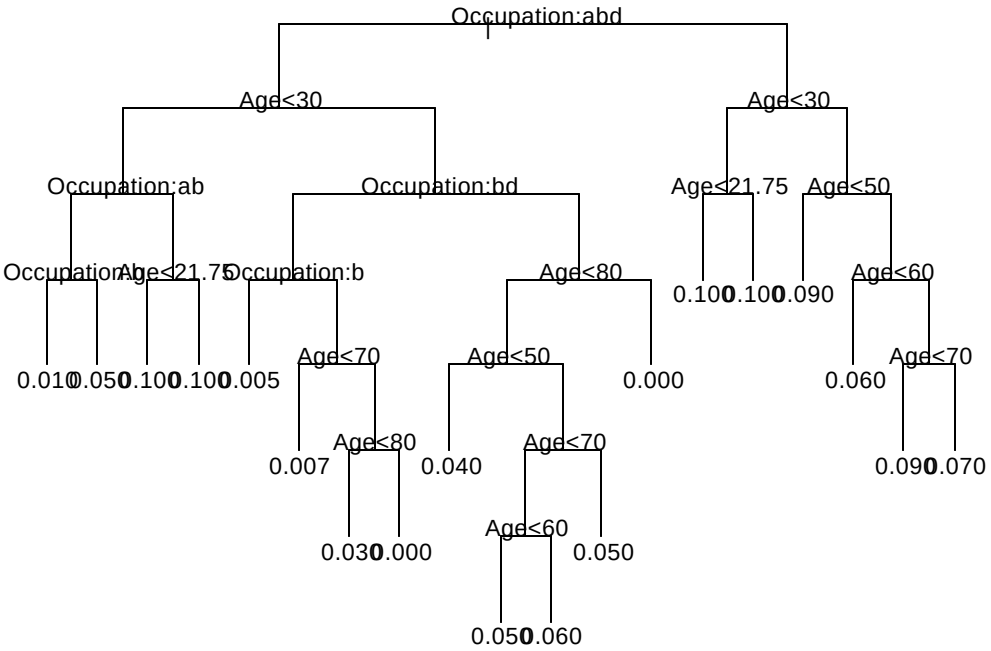
First Split


Base Node Frequency = .045
Occu3
Frequency = .04
Occu1, Occu2, Occu4
Frequency = .09

Next 2 Splits



Final Tree

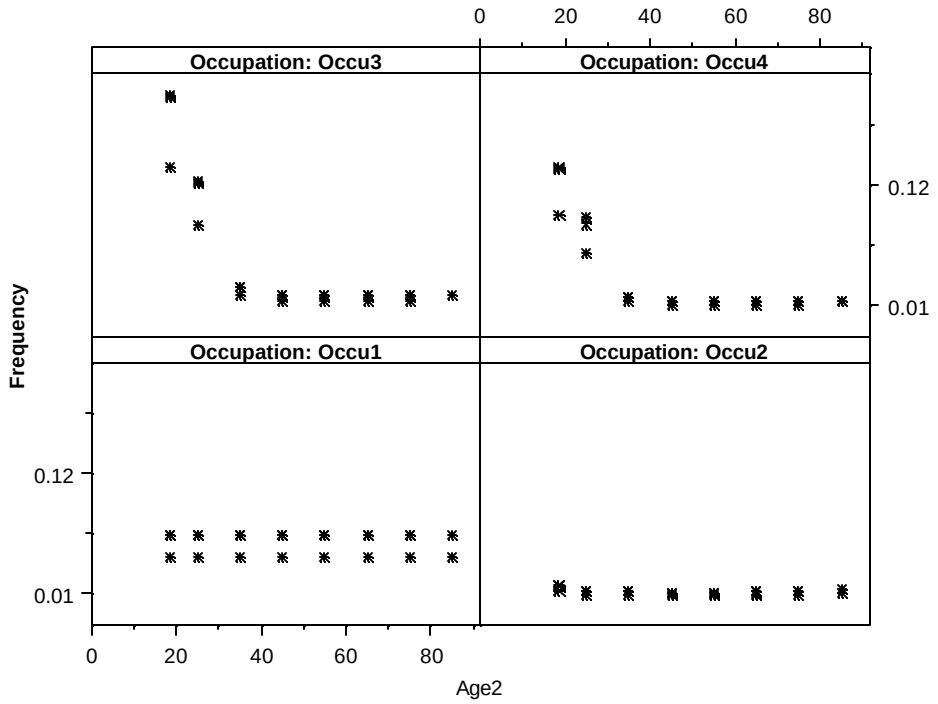

Occupation:abd
Age<30
Age<30
Occupation:ab
Occupation:bd
Age<21.75
Age<50
Occupation:a
Age<21.75
Occupation:b
Age<80
Age<60
0.100
0.100
0.090
Age<70
Age<50
Age<70
0.010
0.050
0.100
0.100
0.005
0.000
0.060
Age<80
Age<70
0.007
0.040
0.090
0.070
Age<60
0.030
0.000
0.050
0.050
0.060

Classification Applications

- Another type of problem:
  - Which of two groups does a new policyholder belong to?
- Example: Suppose all records are classified as high frequency or other
- We want to predict which group a new observation belongs to

Revised Frequency Data:

Add variable company size



Example: Classify High Frequency Observations

Frequency	High/ Low	Co Size	Occu2	Occu3	Occu4	Age<35	Interact Age	Age* occu2	Age* occu3	Age* occu4
0.4%	0	1	1	0	0	0	0	55	0	0
0.0%	0	1	1	0	0	0	0	65	0	0
2.3%	0	1	1	0	0	0	0	75	0	0
0.0%	0	1	1	0	0	0	0	85	0	0
25.0%	1	1	0	1	0	1	18.5	0	18.5	0
10.5%	1	1	0	1	0	1	25	0	25	0
2.6%	0	1	0	1	0	1	35	0	35	0
1.5%	0	1	0	1	0	0	0	0	45	0

Discriminant Analysis

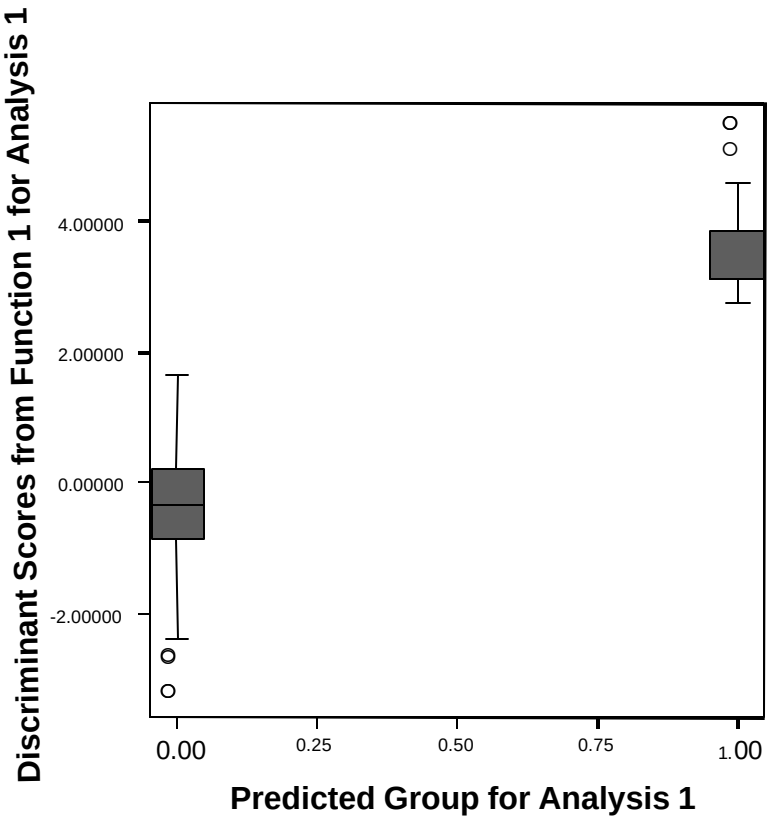
- A classical procedure for classification
- Finds the combination of variables which maximizes the difference between the two groups
- Produces a score from predictor variables. Score is used for classification
- Like a regression with a binary dependent variable

Discriminant Analysis cont.

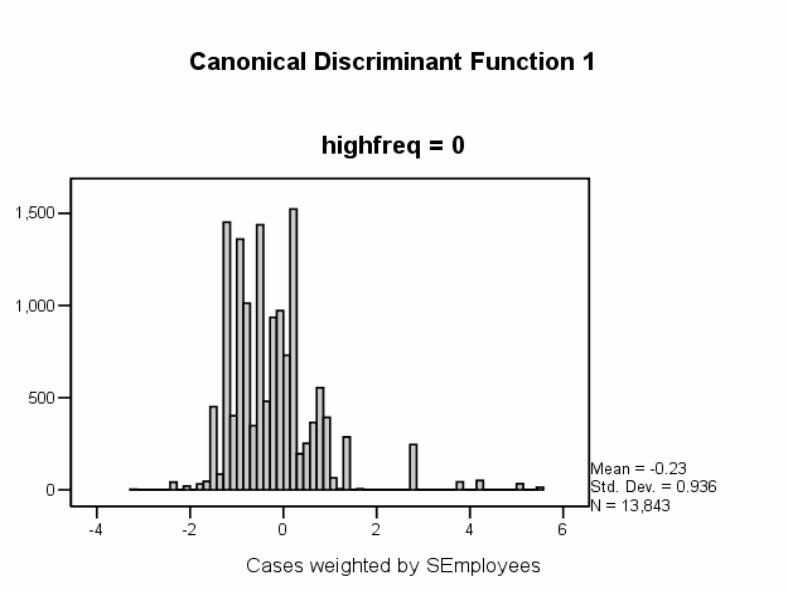
- Similar to a binary regression:
- $z = a + b_1x_1 + b_2x_2 + \ldots b_nx_n$
  - z is a binary dependent variable (i.e., values of 0 and 1)
  - x's are predictor variables

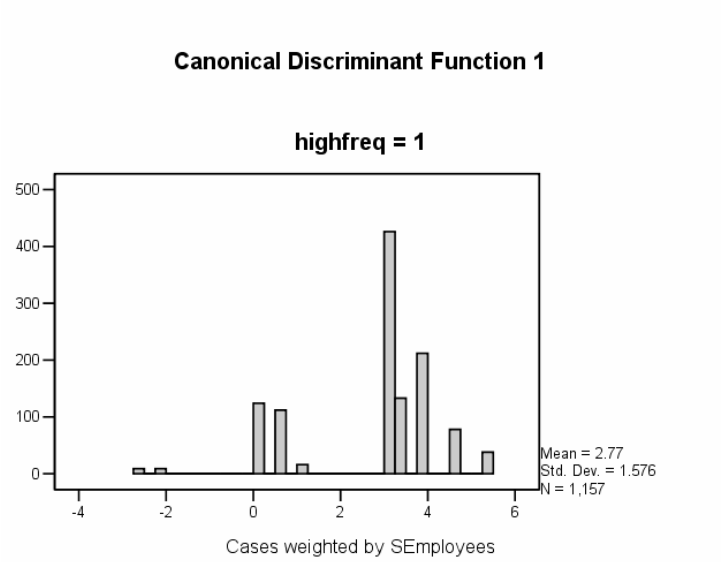
Score Discriminates between the Classes

Box and Whisker Plots of Discriminant Scores by Group



Score Discriminates between the Classes





Discriminant Analysis

Standardized Canonical Discriminant Function Coefficients

	Function 1
Age	*.690*
Size	*.180*
Occu2	*-.109*
Occu3	*2.146*
Occu4	*2.084*
Age<35	*2.745*
Age*(Age< 35)	*-2.376*
Age*occu2	*-.029*
Age*occu3	*-1.826*

Other Methods for Classification

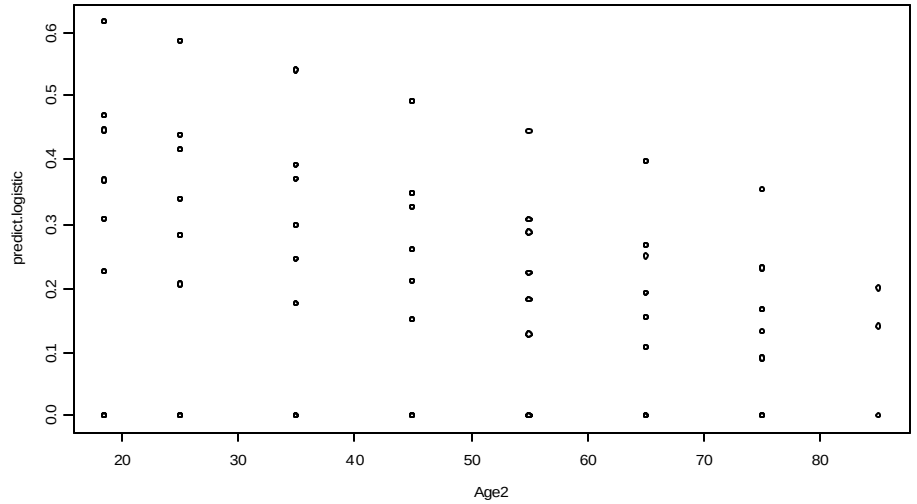
- Logistic regression
  - Can use with GLM's
  - Uses logit link: ln(p/(1-p))
    - This is log of odds ratio

$$\ln\left(\frac{p}{1-p};\mathbf{x}\right) = B_0 + B_1 X_1 + \ldots + B_n X_n$$

  - Assumes binomial distribution

Logistic Regression for Classification

```
# use GLM function to fit logistic regression
freq.glm<-
glm(sumemp$High~sumemp$Age2+sumemp$Occupation+sumemp$Miles,family=binomial,data=sumemp
)
# get predicted value
predict.logistic<-predict.glm(freq.glm,type=c(”response”),ci.fit=T)
plot(Age2,predict.logistic)
```



Other Methods for Classification cont.

- Decision Trees/CART
  - A classification version of trees with binary dependent variable.
  - Typically a different goodness of fit measure than R^2.
- Support Vector Machines

Which Model to Use?

- Regression with normality assumption is one of GLM options
  - Most insurance distributions are skewed
  - Might want to use GLM with Gamma distribution
- However, if normality is a reasonable assumption, many more diagnostic tools are available for regression
  - Lognormal – when data is logged it is normal
- The distributional forms of much of insurance data is more heavy-tailed than any exponential family member
  - Robust methods

Which Model to Use?

- The link functions do not capture all the possible non-linear relationships
  - Non-parametric methods such as regression trees
  - Kernel regression
- Can use more than one model
  - This is usually recommended

Which Model to Use?

- Classification
  - Discriminant analysis is a reasonable method to start out with
  - Logistic regression is more frequently used than discriminant analysis. Its use requires special software (such as R)
  - To capture data complexities such as nonlinearities, trees are easy to understand, common method

Introductory Modeling Library Recommendations

- Berry, W., *Understanding Regression Assumptions*, Sage University Press
- Iversen, R. and Norpoth, H., *Analysis of Variance*, Sage University Press
- Fox, J., *Regression Diagnostics*, Sage University Press
- Chatfield, C., *The Analysis of Time Series*, Chapman and Hall
- Fox, J., *An R and S-PLUS Companion to Applied Regression*, Sage Publications
- *2004 Casualty Actuarial Discussion Paper Program* on Generalized Linear Models, www.casact.org

Advanced Modeling Library Recommendations

- Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997
- Francis, Louise, 2003, “Martian Chronicles: Is MARS Better than Neural Networks” at www.casact.org/aboutcas/mdiprize.htm
- Francis, Louise, 2001, “Neural Networks Demystified” at www.casact.org/aboutcas/mdiprize.htm