Understanding the Essential Tools and Techniques of Data Mining

Prepared by
Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com
Louise.francis@data-mines.com
March 27, 2006

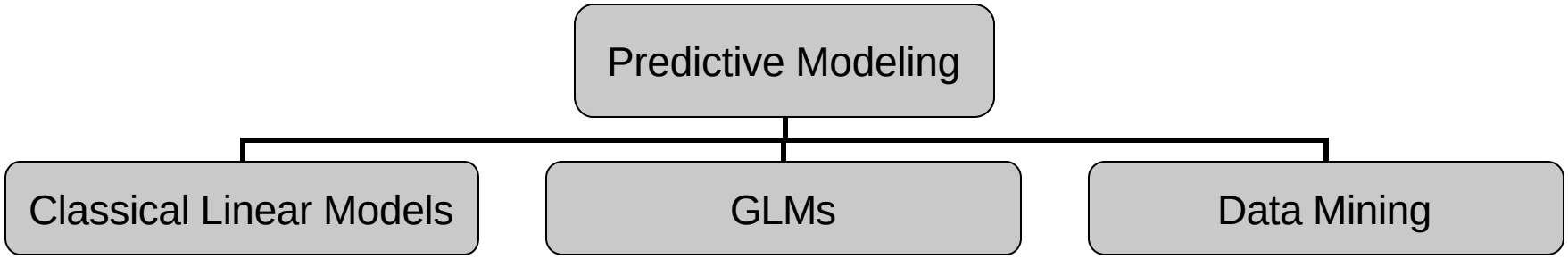
Modeling 200 Objectives

- Gentle introduction to some advanced classical statistical models and
- Introduction to intelligent methods for pattern recognition
- Illustrate with some simple workers compensation applications
- Show examples in commonly available software (see Excel files that accompany slides)
- Discuss practical modeling issues

This is a continuation of Intro to Predictive Modeling

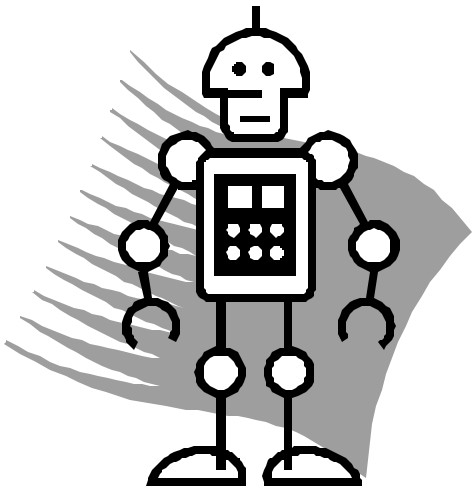
- For refresher see materials on www.data-mines .com
- Focus was on linear regression and non-linear regression

Predictive Modeling Family



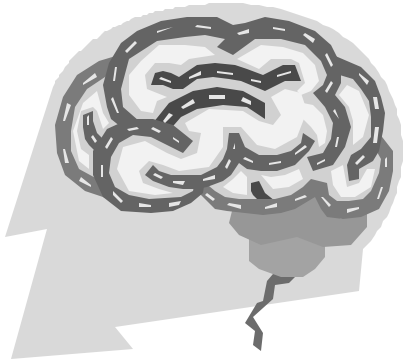
Why Data Mining?

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



Frontiers of Modeling

- Will cover some of latest analytic methods
  - Naïve bayes
  - Logic/decision trees
  - Neural networks



Kinds of Applications

- Classification
  - Categorical dependent variable
    - Fraud detection
    - Prospective identification of catastrophic claims
- Prediction
  - Numeric dependent variable
    - Frequency
    - Severity
    - Loss ratio
    - Profitability

The Data

- Closed claim data obtained from Texas Department of Insurance Web Site
- Only larger claims
- Selected only employment related claims
- Dependent variables:
  - Claim severity (Paid Loss)
  - Serious (>$300,000)/Non-Serious Indicator

The Data cont.

- Predictor Variables
  - Injury
  - Cause of Loss
  - Age of claimant
  - Attorney involvement
  - Health insurance
  - Many others
    - Economic damages
    - Punitive damages
    - Initial reserves

Categorical Data: Contingency Table Analysis

- Create Crosstabulations of data
- Two way tabulation of counts in cells

Burns (heat) * Serious Crosstabulation

Count

		Serious		
		.00	1.00	Total
Burns (heat)	N	1400	343	1743
	Y	41	34	75
Total		1441	377	1818

Crosstabulations

- 9% of burns are serious claims vs. 2.8% of non- burns
- Is there a correlation between burn and serious?

Burns (heat) * Serious Crosstabulation

			Serious		
			.00	1.00	Total
Burns (heat)	N	Count	1400	343	1743
		% within Serious	97.2%	91.0%	95.9%
	Y	Count	41	34	75
		% within Serious	2.8%	9.0%	4.1%
Total		Count	1441	377	1818
		% within Serious	100.0%	100.0%	100.0%

Chi Square Statistic

$$c^2 = \sum \frac{(O-E)^2}{E}$$

O = Observed (Actual) cell count

E=Expected Cell Count (under independence)

= row percent* column percent * Total count

Degrees of Freedom

- For regression and ANOVA: “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*
- For Chi Square counts rather than sums matter
- We need DF to test significance of relationship
  - We look up the Chi-Square value for a given degrees of freedom

Degrees of Freedom for Chi Square

- There are four cells in total
- Formula (rows – 1)*(columns – 1)
- This is (2 - 1) * (2 – 1) =1
- We know the row percent serious and the column percent serious, and the total count which leaves only one free parameter

Chi Square Statistic - Burns

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	28.792[b]	1	.000		
Continuity Correction[a]	27.253	1	.000		
Likelihood Ratio	23.923	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	28.777	1	.000		
N of Valid Cases	1818				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15. 55.

Correlation Measure: Contingency Coefficient

- Categorical variable correlation or measure of association
- Varies between 0 and 1
- 0 = no association, 1 = perfect association

$$p = \sum_i \sum_j \frac{p_{ij} - p_i p_j}{p_i p_j}$$

Contingency Coefficient - Burns

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.125	.000
N of Valid Cases		1818	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Chi Square and Contingency Coefficient – All Injuries

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	205.876[a]	16	.000
Likelihood Ratio	201.862	16	.000
Linear-by-Linear Association	6.028	1	.014
N of Valid Cases	1818		

a. 12 cells (35.3%) have expected count less than 5. The minimum expected count is .41.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.319	.000
N of Valid Cases		1818	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

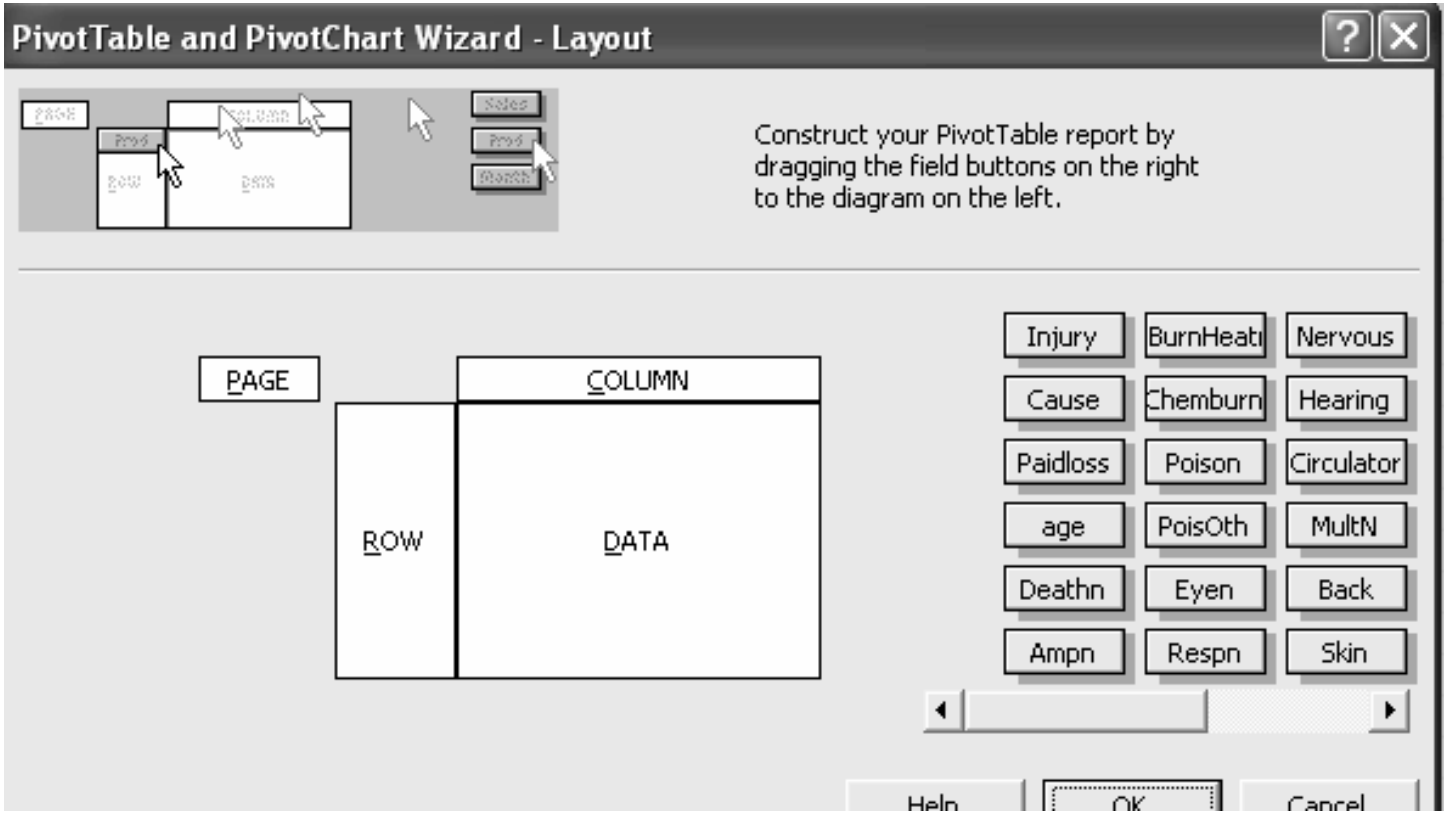
Crosstabulations in Excel: Pivot Tables

- One of easiest ways in Excel to get cross tabulations is to use Pivot Tables
- Highlight (i.e, select) range where data is
- On Data toolbar
  - Click pivot table

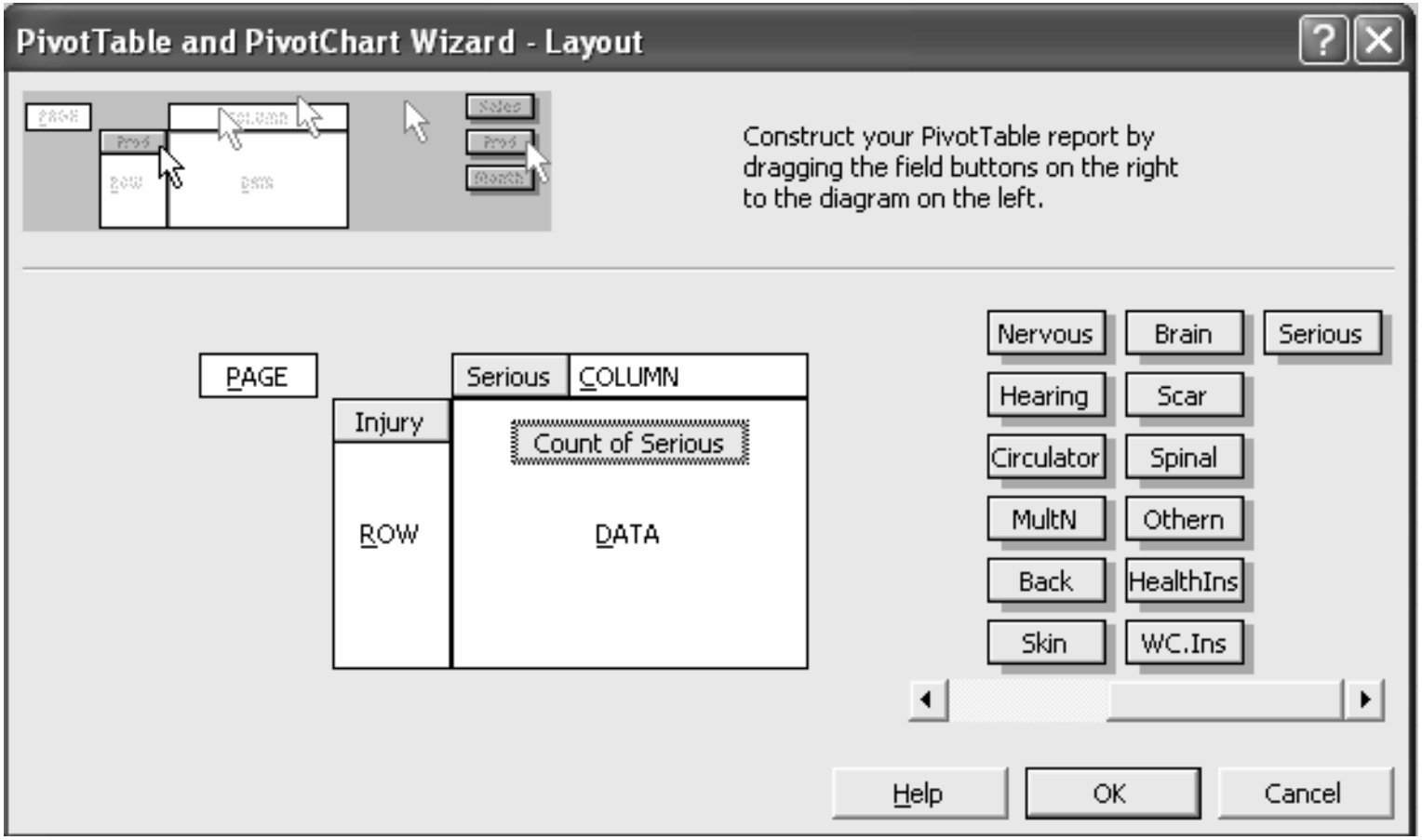
Pivot Table: Specify Range

Pivot Table: Specify where output goes

Pivot Table- Specify Layout



Pivot Table – Specify Injury as Row and Serious as Column and count of Serious as statistic



Pivot Table: Click OK, Finish

Count of Serious	Serious		
Injury	0	1	Grand Total
1	28	11	39
2	26	19	45
3	15	6	21
4	16		16
6	5		5
7	37		37
8	4		4
9	4	2	6
10	1	1	2
11	492	129	621
12	410	52	462
13	9		9
14	15	11	26
15	12	8	20
16	15	11	26
17	253	27	280
18	99	100	199
Grand Total	1441	377	1818

Pivot Table-Assign Labels to Injury

Injury	Injury Code	Non Serious	Serious	Grand Total
Amputation	1	28	11	39
Burnsheat	2	26	19	45
Burnschemical	3	15	6	21
Systemicpoisoningtoxic	4	16		16
Systemicpoisoningother	5	0	0	0
Eyeinjuryblindness	6	5		5
Respiratorycondition	7	37		37
Nervouscondition	8	4		4
Hearinglossorimpairment	9	4	2	6
Circulatorycondition	10	1	1	2
Multipleinjuries	11	492	129	621
Backinjury	12	410	52	462
Skindisorder	13	9		9
Braindamage	14	15	11	26
Scarring	15	12	8	20
Spinalcordinjuries	16	15	11	26
Other	17	253	27	280
Death	18	99	100	199
	Grand Total	1463	355	1818

Use Pivot Table to Compute Chi-Square

Data for Calculation

	Row Percent (serious)	Column Percent (Non-Serious)	Column Percent (Serious)
Injury	**Cell Count/Row Total**	**Non Serious/Total**	**Serious/Total**
Amputation	2.1%	80.5%	19.5%
Burnsheat	2.5%	80.5%	19.5%
Burnschemical	1.2%	80.5%	19.5%
Systemicpoisoningtoxic	0.9%	80.5%	19.5%
Systemicpoisoningother	0.0%	80.5%	19.5%
Eyeinjuryblindness	0.3%	80.5%	19.5%
Respiratorycondition	2.0%	80.5%	19.5%
Nervouscondition	0.2%	80.5%	19.5%
Hearinglossorimpairment	0.3%	80.5%	19.5%
Circulatorycondition	0.1%	80.5%	19.5%
Multipleinjuries	34.2%	80.5%	19.5%
Backinjury	25.4%	80.5%	19.5%
Skindisorder	0.5%	80.5%	19.5%
Braindamage	1.4%	80.5%	19.5%
Scarring	1.1%	80.5%	19.5%
Spinalcordinjuries	1.4%	80.5%	19.5%
Other	15.4%	80.5%	19.5%
Death	10.9%	80.5%	19.5%

Compute Chi Square Statistic

Expected		Chi Square Statistic ((Observed-Expected)^2)/Expected	
Non-Serious	Serious		
row* non-serious* N	row* Serious*N	Non-Serious	Serious
31.4	7.6	0.4	1.5
36.2	8.8	2.9	11.9
16.9	4.1	0.2	0.9
12.9	3.1	0.8	3.1
-	-	-	-
4.0	1.0	0.2	1.0
29.8	7.2	1.8	7.2
3.2	0.8	0.2	0.8
4.8	1.2	0.1	0.6
1.6	0.4	0.2	1.0
499.7	121.3	0.1	0.5
371.8	90.2	3.9	16.2
7.2	1.8	0.4	1.8
20.9	5.1	1.7	6.9
16.1	3.9	1.0	4.3
20.9	5.1	1.7	6.9
225.3	54.7	3.4	14.0
160.1	38.9	23.3	96.2

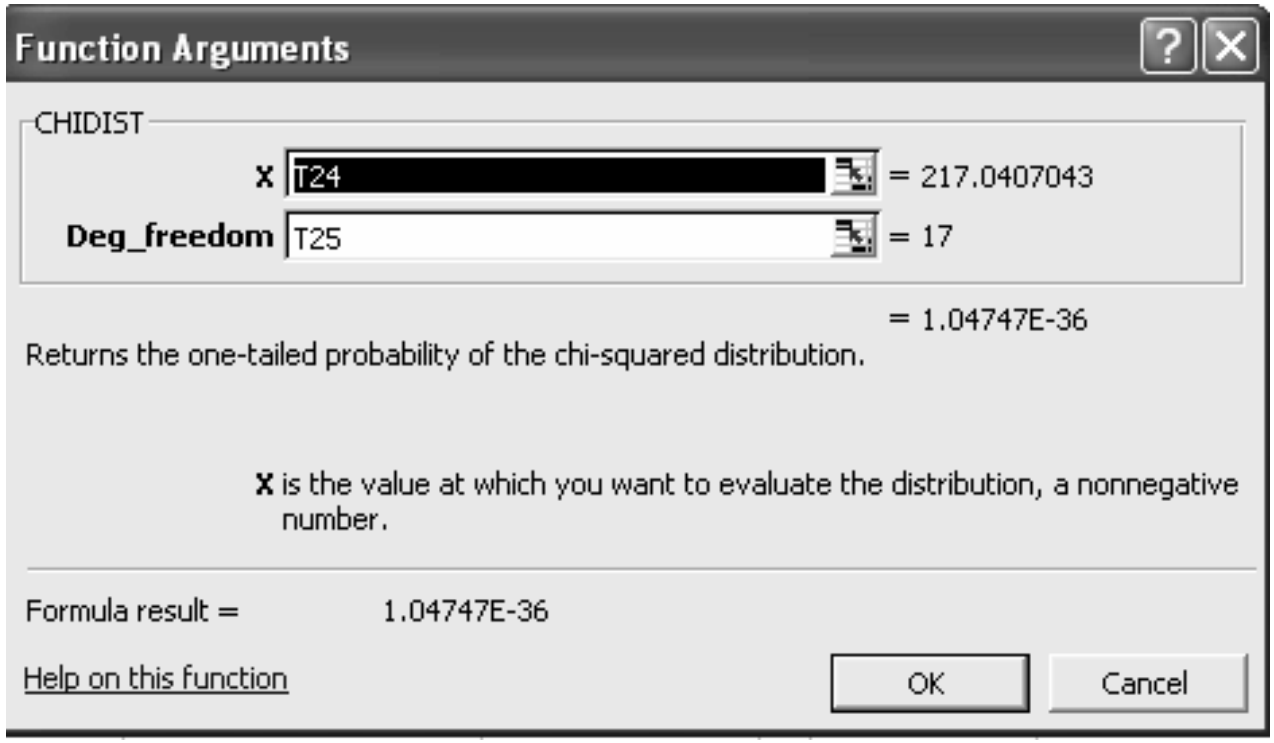
Sum Chi-Square	217.0
df = (r-1)(C-1)	17
Chi Square p	1.04747E-36

Result: Statistically Significant

Conclusion: Significant association

Use Chi Square Distribution Function to Test Significance

- Specify Chi Square value and df, functions gives probability
- Probability < 5% generally considered significant

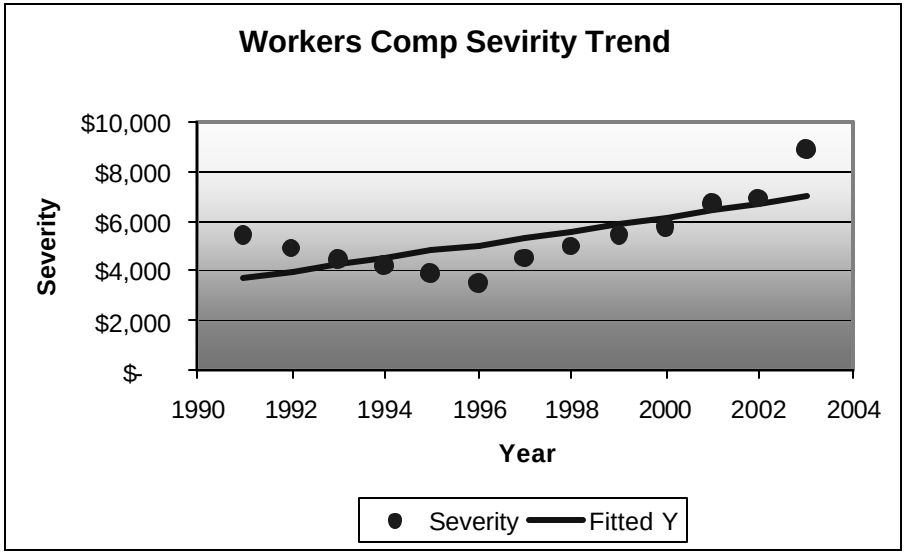


Hypothesis Test vs Prediction

- Chi square typically used to test significance of an association
- We are interested in prediction, not hypothesis testing
- Chi square is used in one of the common data mining models: decision trees
  - We will return to it to see how it is used
- We now move to other methods to predict categorical variables
  - Naïve bayes
  - Logistic regression

Review from Last Workshop: Regression for Prediction

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



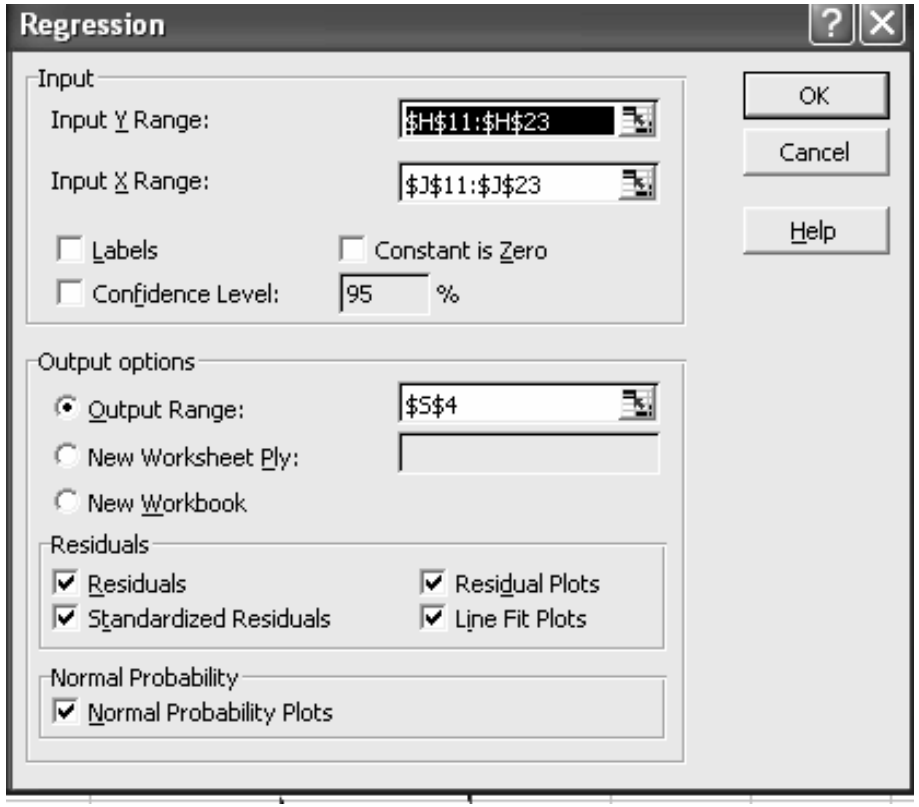
A Brief Review of Regression

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

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

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			

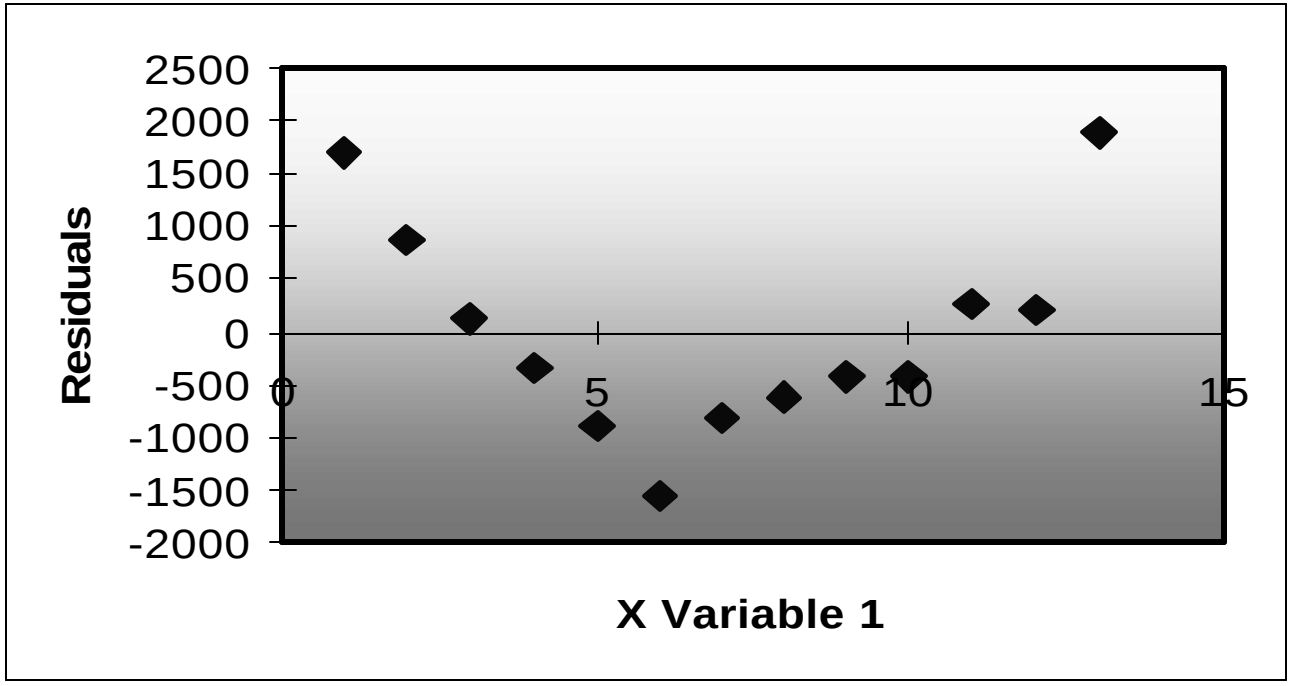
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3445.41	619.37	5.56	0.00	2082.18	4808.64
X Variable 1	270.02	78.03	3.46	0.01	98.27	441.77

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



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)

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

Logistic Regression

- A very common method for modeling categorical dependent variables
- A member of the GLM family of models
  - Distribution typically assumed to be binomial

$$P(X = x, N) = \binom{N}{x} p^x (1-p)^{N-x}$$

  - For instance the probability of three heads in four tosses of a coin is:

$$P(X = 3, N = 4) = \binom{4}{3} \frac{1}{2}^3 (1-\frac{1}{2})^{4-3} = \frac{4!}{3!1!} \frac{1}{2}^4 = 0.0156$$

Odds Ratio

- The odds ratio is the ratio of the probability of success to the probability of non-success

$$OddsRatio = \frac{p}{1-p}$$

p = probability of success

i.e: probability of serious claim

1-p=probability of non-serious claim

- We can compute this in Excel from our Texas data

Odds and Log Odds for Injury Data

Actual

Injury	Non Serious	Serious
Death	99	100
Back Injury	410	52
	509	152
Column Percent	253%	76%

ODDS Ratio: Ratio of srerious to non-serious

Injury	Odds Ratio	Log Odds
Death	1.010	0.0100503
Back Injury	0.127	-2.064913

Implementation in Excel

- Dependent variable is log of odds ratio or logit (hence logistic regression)
- Independent variables are:
  - Age group: use the midpoint of the age buckets
  - 14-30, 31-37, 38-43, 44-50, 51-83
  - Brain-spinal injury (a brain or spinal cord injury)
  - Death
  - Burns
- We drop assumption of binomial distribution
- Use Regression procedure in Excel

Logistic Regression: Practical considerations

- Use Aggregated data
- The Odds ratio is undefined if non-serious=0, so test for 0 and add small constant (0.01)
- The log odds is undefined if serious is 0, so add small constant (such as 0.01)

Output From “Logistic” Regression”

Full model: logodds ~ age + brainspinal + death + burns

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.64984591
R Square	0.4222997
Adjusted R Sq	0.31726328
Standard Erro	0.34260649
Observations	27

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**4**	**1.887696**	0.471924	4.020507	**0.013502534**
Residual	**22**	**2.582342**	0.117379		
Total	26	4.470038			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.2016383	0.219018	-0.92065	**0.367221**	-0.65585	0.25258
agegroupavg	-0.00788884	0.00433	-1.8218	**0.082111**	-0.01687	0.00109
brainspinaln	0.4165116	0.150089	2.775092	**0.011042**	0.10525	0.72778
deathn	0.36856769	0.136655	2.697059	**0.013165**	0.08516	0.65197
burnsn	0.1368833	0.150152	0.911632	**0.371844**	-0.17451	0.44828

Assessment of Logistic Regression: Full Model

- Not a very good fit
- Burns is least significant predictor

Reduce Size of Model

Reduced model: logodds ~ age + brainspinal + death

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.63283212
R Square	0.40047649
Adjusted R Sq	0.32227777
Standard Erro	0.341346
Observations	27

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**3**	**1.790145**	0.596715	5.121267	**0.007356902**
Residual	**23**	**2.679893**	0.116517		
Total	26	4.470038			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.12770738	0.202702	-0.63002	**0.534888**	-0.54702909	0.29161434
agegroupavg	-0.00800394	0.004312	-1.85599	**0.076313**	-0.01692498	0.00091711
brainspinaln	0.36881092	0.140156	2.631439	**0.01492**	0.078876924	0.65874492
deathn	0.34985393	0.134608	2.599062	**0.016044**	0.071396585	0.62831128

Reduce Model cont.

Can we eliminate the variable **burnsn**?			
Compare full model with reduced model.			
	Full model	Reduced Model	
Residual sum of squares (RSS)	2.5823	2.6799	
	p =	q =	
Number variables in model	5	4	
n= number observations	27		
Numerator of F statistic for comparison =			
Difference in RSS / (p-q)	0.0976		
Denominator of F statistic			
RSS full model / (n-p)	0.1174		
F-value = Numerator/Denominator	**0.8311**		
Parameters for F: p-q,n-p (	1	22)	
The 5% confidence value for F(1,22) is	**4.3**		
Since the F value is less than the critical value,			
we can drop the variable **burnsn** from the model.			

Reducing the size of a Model: Comments

- The model design used was not ideal:
- Too few elements in some of the summarized records.
- Some cells had one claim only.
- No recognition of relative weight by cell.

Next attempt:

- Summarize data further.
- Reduce number of variables.

Summarize Data Further

Observed summarized data							
brainspinal	agegroup2	agegroup	agegroup avg	seriousn	nonseriousn	empirical p	logodds
0	14-50	14-30	21	81	266	0.2334	-1.1890
0	14-50	31-37	34	75	288	0.2066	-1.3455
0	14-50	38-43	40	75	295	0.2027	-1.3695
0	14-50	44-50	47	61	249	0.1968	-1.4066
0	50-83	50-83	65	64	278	0.1871	-1.4687
1	14-50	14-30	21	9	9	0.5000	0.0000
1	14-50	31-37	34	6	4	0.6000	0.4055
1	14-50	38-43	40	10	7	0.5882	0.3567
1	14-50	44-50	47	15	10	0.6000	0.4055
1	50-83	50-83	65	5	11	0.3125	-0.7885
				401	1417		

Refining the Fit

- Logodds is the dependent variable.
- Use average age in each bucket as numerical predictor
- Use brain/spinal cord injury (br-sp) as binary categorical predictor
- Allow for interaction between age and br-sp.
- Introduce “artificial” variable to segregate 51-83 category. 51-83 category with br-sp injury is an outlier.
- Detailed analysis of fitting process provided by Joe Marker, FCAS, MAAA

Revised Model

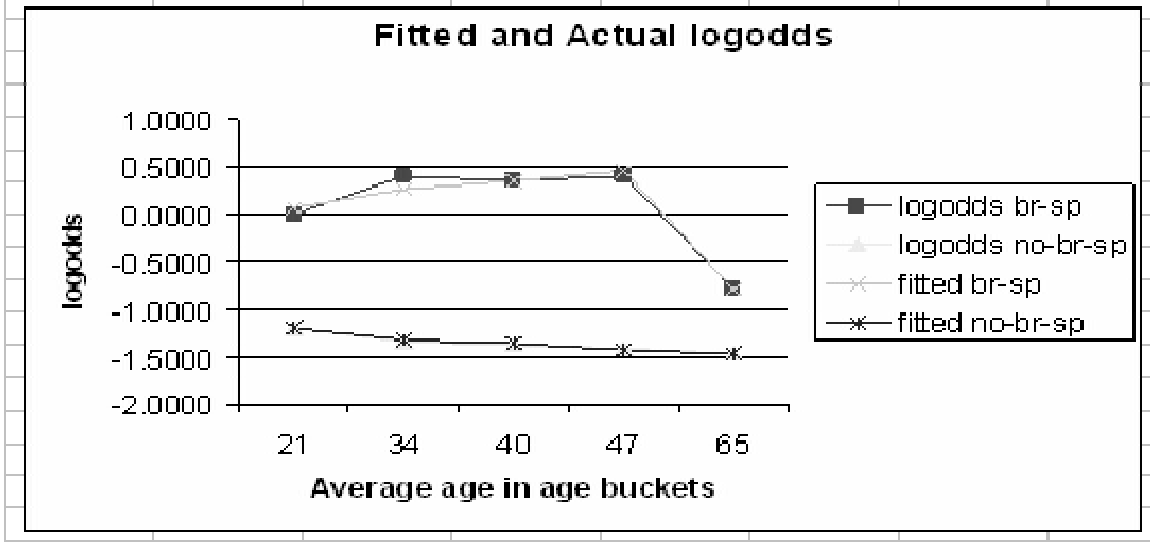
SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.997667					
R Square	0.99534					
Adjusted R Squ	0.989514					
Standard Error	0.085106					
Observations	10					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	5	6.187769	1.237554	170.8601	9.45785E-05	
Residual	4	0.028972	0.007243			
Total	9	6.216741				
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-1.02819	0.163766	-6.27839	0.003286	-1.48287288	-0.5735
brains	0.769624	0.2316	3.323078	0.029294	0.126600193	1.412649
ageavg	-0.00844	0.004455	-1.89363	0.131199	-0.02080364	0.003933
group2ind	0.107756	0.162244	0.664159	0.542917	-0.34270601	0.558218
brainsxgroup2in	-1.64554	0.229448	-7.17175	0.002002	-2.28259053	-1.00849
brainxavg	0.023942	0.0063	3.800328	0.019098	0.00645031	0.041433

Final Model: Actual vs Fitted Value

Model: logodds ~ average age by brain-spinal indicator

ageavg	logodds br-sp	logodds no-br-sp	fitted br-sp	fitted no-br-sp
21	0.0000	-1.1890	0.0671	-1.2053
34	0.4055	-1.3455	0.2686	-1.3150
40	0.3567	-1.3695	0.3617	-1.3656
47	0.4055	-1.4066	0.4702	-1.4247
65	-0.7885	-1.4687	-0.7885	-1.4687

Note: Age 51-83 separated out as a separate category because its logodds for brain-spinal category is an outlier.



br-sp means brain/spinal cord injury

no-br-sp means no brain/spinal cord injury

R/S-Plus Code

Naive Bayes

- Naive Bayes assumes conditional independence
- Probability that an observation will have a specific set of values for the independent variables is the product of the conditional probabilities of observing each of the values given category cj

$$P(X \mid c_j) = \prod_j P(x_i \mid c_j)$$

Naïve Bayes

$$P(C_i \mid X_1, X_2 \ldots X_N) = \frac{p(C_i)\prod_j P(x_i \mid c_j)}{\prod_j P(x_i)}$$

Naïve Bayes

- Only classification problems (categorical dependents)
- Only categorical dependents
- Create groupings for numeric variables such as age
  - We split age into quintiles (5 groups)

Naïve Bayes – Starting Point is Pivot Table

Count of Serious	Serious		
Cause	0	1	Grand Total
Airtransportation	1		1
Drowning	4		4
Explosions	16	17	33
Falls	351	76	427
Fire	9	7	16
Firearm	5	2	7
Offroadvehicle	28	13	41
Oilgasextraction	27	16	43
Other_A	276	92	368
Othermotorvehicl	554	121	675
PollutionToxicexp	58	1	59
Railway	8	3	11
Surgicalmedicalca	27	6	33
Useofagriculturaln	6		6
Useofdefectivepro	71	23	94
Grand Total	1441	377	1818

% Serious Given Independent Variable
0.0%
0.0%
51.5%
17.8%
43.8%
28.6%
31.7%
37.2%
25.0%
17.9%
1.7%
27.3%
18.2%
0.0%
24.5%
20.7%

Naïve Bayes – Starting Point is Pivot Table cont.

Count of Serious	Serious		
Age Group	0	1	Grand Total
1	245	76	321
2	291	75	366
3	258	59	317
4	311	93	404
5	336	74	410
Grand Total	1441	377	1818

% Serious Given Independent Variable
23.7%
20.5%
18.6%
23.0%
18.0%

Naïve Bayes: Probability of Independent Variable Given Serious/Non Serious

Count of Serious	Serious		
Cause	0	1	Grand Total
Airtransportation	1		1
Drowning	4		
Explosions	16		
Falls	351		
Fire	9	7	16
Firearm	5	2	7
Offroadvehicle	28	13	41
Oilgasextraction	27	16	43
Other_A	276	92	368
Othermotorvehicle	554	121	675
PollutionToxicexp	58	1	59
Railway	8	3	11
Surgicalmedicalca	27	6	33
Useofagriculturaln	6		6
Useofdefectivepro	71	23	94
Grand Total	1441	377	1818

Probability Cause Given Non Serious
0.0007
0.0028
0.0111
0.2436
0.0062
0.0035
0.0194
0.0187
0.1915
0.3845
0.0402
0.0056
0.0187
0.0042
0.0493
1.0000

divide cell by row total for non serious claims

Naïve Bayes: Probability of Independent Variable Given Serious/Non Serious cont.

Count of Serious	Serious		
Age Group	0	1	Grand Total
1	245	76	321
2	291	75	366
3	258	59	317
4	311	93	404
5	336	74	410
Grand Total	1441	377	1818

Probability Age Given Non Serious	Probability Age Given Serious
0.1700	0.2016
0.2019	0.1989
0.1790	0.1565
0.2158	0.2467
0.2332	0.1963
1.0000	1.0000

Put it together to make a prediction

- Plug in specific values for Injury, cause and age group and get a prediction
- Chain together product of injury* probability of cause * probability of age * probability non-serious

Inj Code	Cause	Age Group	Probability Injury	Probability Cause	Probability Age	Probability Injury
1	Useofdefective	2	0.0194	0.0493	0.2019	0.0292
1	Other_A	2	0.0194	0.1915	0.2019	0.0292

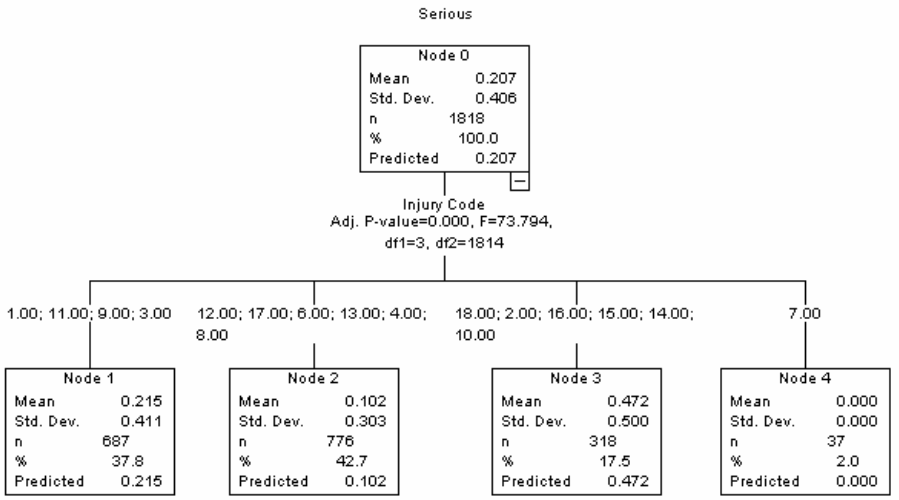
Probability Cause	Probability Age	Product-NS	Product-S	Ratio (NS/S)	Class
0.0610	0.1989	0.0002	7.34E-05	2.09E+00	Non-Serious
0.2440	0.1989	0.0006	2.94E-04	2.03E+00	Non-Serious

Advantages/Disadvantages

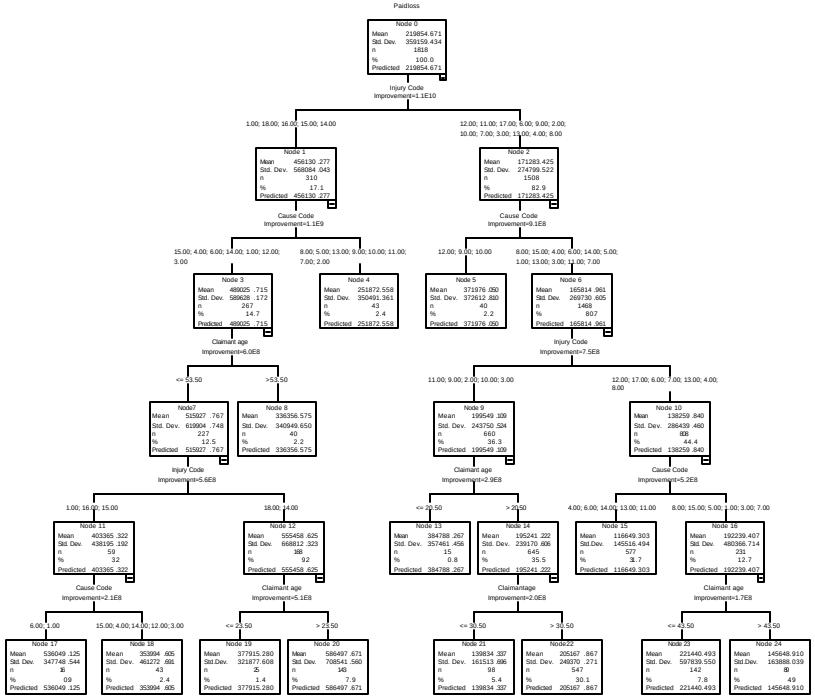
- Computationally efficient
- Under many circumstances has performed well
- Assumption of conditional independence often does not hold
- Can’t be used for numeric variables

Trees – One of the most common methods

- Creates a set of logic rules bases on recursive partitioning of the data



Tree for Numeric Dependent



Trees – Example of Tree Logic Rules

```
/* Node 1 */
DO IF (VALUE(injury) EQ 1 OR VALUE(injury) EQ 11 OR VALUE(injury) EQ 9
OR VALUE(injury) EQ 3).
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 0.215429403202329.
END IF.
EXECUTE.

/* Node 2 */
DO IF (SYSMIS(injury) OR VALUE(injury) NE 1  AND  VALUE(injury) NE 11
AND  VALUE(injury) NE 18  AND  VALUE(injury) NE 9  AND  VALUE(injury)
NE 2  AND  VALUE(injury) NE 16  AND  VALUE(injury) NE 15  AND
VALUE(injury) NE 14  AND  VALUE(injury) NE 10
  AND  VALUE(injury) NE 7  AND  VALUE(injury) NE 3).
COMPUTE nod_001 = 2.
COMPUTE pre_001 = 0.10180412371134.
END IF.
EXECUTE.

/* Node 3 */
DO IF (VALUE(injury) EQ 18 OR VALUE(injury) EQ 2 OR VALUE(injury) EQ 16
OR VALUE(injury) EQ 15 OR VALUE(injury) EQ 14 OR VALUE(injury) EQ 10).
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 0.471698113207547.
END IF.
EXECUTE.

/* Node 4 */
DO IF (VALUE(injury) EQ 7).
COMPUTE nod_001 = 4.
COMPUTE pre_001 = 0.
END IF.
EXECUTE.
```

One of Earliest Methods - CHAID

- Chi square automatic interaction detection
- Uses Chi square goodness of fit statistic to build model
- First merge together categories that are not statistically different

Combine categories

				Actaul Column %	
Injury	**Non Serious**	**Serious**	**Grand Total**	**Non-Serious %**	**Serious %**
Death	99	100	199	49.7%	50.3%
Circulatorycondition	1	1	2	50.0%	50.0%
Braindamage	15		26	57.7%	42.3%
Spinalcordinjuries	15		26	57.7%	42.3%
Burnsheat	26		45	57.8%	42.2%
Scarring	12	8	20	60.0%	40.0%
Hearinglossorimpairment	4	2	6	66.7%	33.3%
Burnschemical	15	6	21	71.4%	28.6%
Amputation	28	11	39	71.8%	28.2%
Multipleinjuries	492	129	621	79.2%	20.8%
Backinjury	410	52	462	88.7%	11.3%
Other	253	27	280	90.4%	9.6%
Systemicpoisoningtoxic	16		16	100.0%	0.0%
Systemicpoisoningother	0	0	0	100.0%	0.0%
Eyeinjuryblindness	5		5	100.0%	0.0%
Respiratorycondition	37		37	100.0%	0.0%
Nervouscondition	4		4	100.0%	0.0%
Skindisorder	9		9	100.0%	0.0%

Combine together cells with similar percents

Use Chi Square statistic to Combine Categories of Predictor Variables

Actual

Injury	Non Serious	Serious	Grand Total	Row Percent
Death	99	100	199	99.0%
Circulatorycondition	1	1	2	1.0%
	100	101	201	
Column Percent	50%	50%		

Expected

Injury	Non Serious	Serious
Death	99.00	100.00
Circulatorycondition	1.00	1.00

Chi Square Statistic

Injury	Non Serious	Serious
Death	0.00	0.00
Circulatorycondition	0.00	0.00
Sum	0.00	
df	1	
Probability	0.994358	

Conclusion: not significantly different

Combine Categories

- Continue to combine categories until only those that are significantly different in their predictions are left
- For numeric variables, only contiguous categories are combined
- For categorical variables, all permutations tested
- Do this for all predictor variables

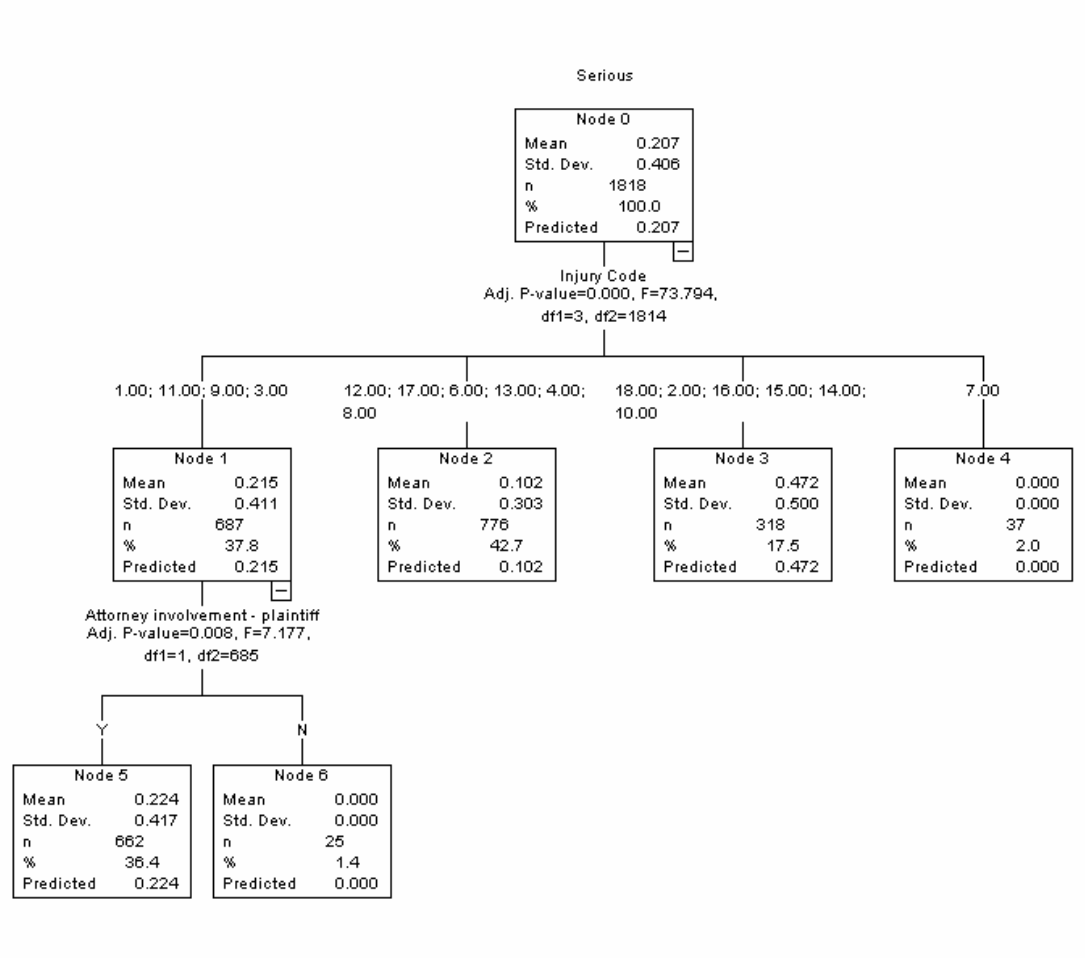
Select variable to split on

- After combining similar categories
  - Compute Chi square statistic for each of the crosstabulations of serious by predictor variable
  - Determine which variable gives the highest Chi square
  - Select that variable to split on in the first step

Repeat the process

- For each node created
- For each predictor variable
  - Combine similar categories
  - Pick best variable to split on
  - Split again at that node
- Keep splitting until no more significant splits can be found

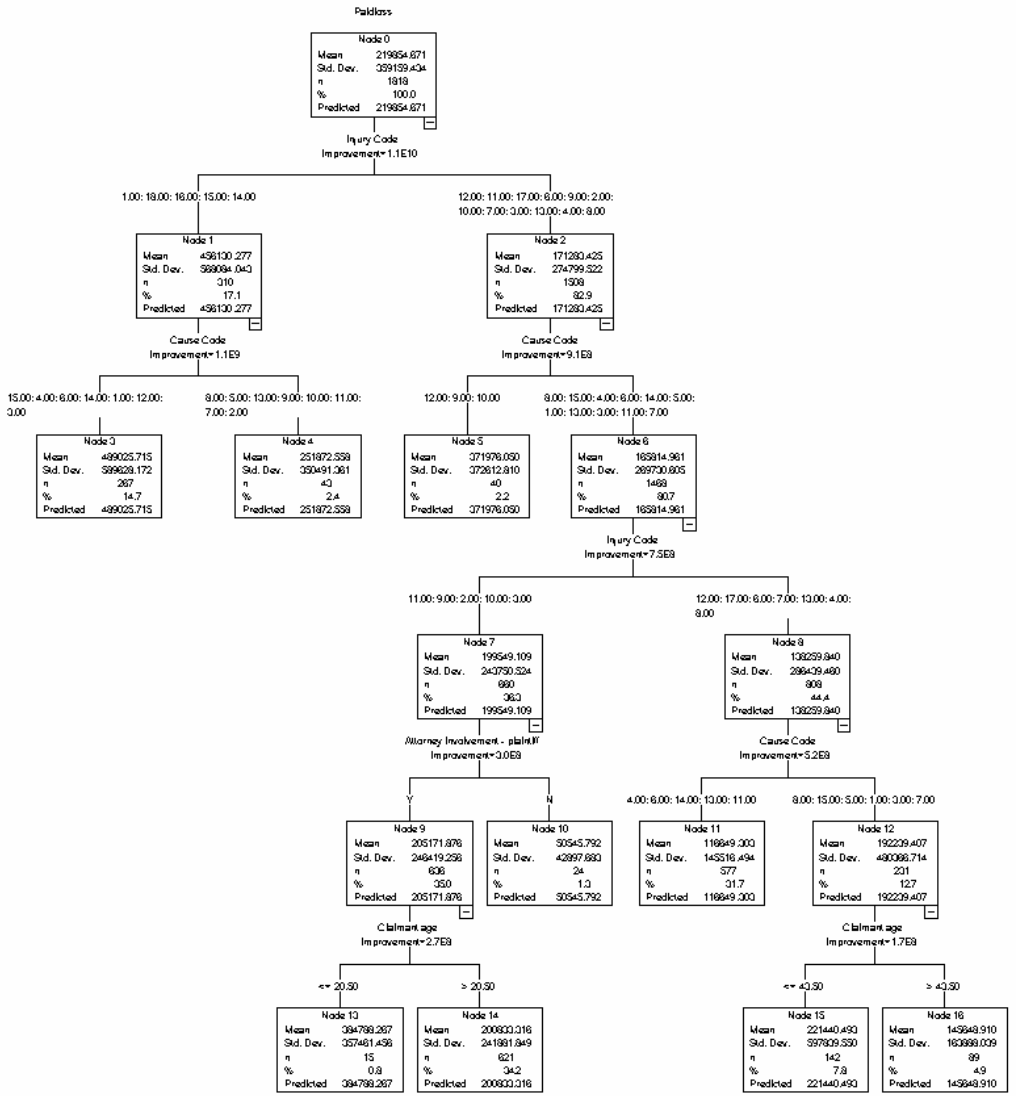
A Final Tree


Serious
Node 0
Mean 0.207
Std. Dev. 0.406
n 1818
% 100.0
Predicted 0.207
Injury Code
Adj. P-value=0.000, F=73.794,
df1=3, df2=1814
1.00; 11.00; 9.00; 3.00
12.00; 17.00; 6.00; 13.00; 4.00; 8.00
18.00; 2.00; 16.00; 15.00; 14.00; 10.00
7.00
Node 1
Mean 0.215
Std. Dev. 0.411
n 687
% 37.8
Predicted 0.215
Node 2
Mean 0.102
Std. Dev. 0.303
n 776
% 42.7
Predicted 0.102
Node 3
Mean 0.472
Std. Dev. 0.500
n 318
% 17.5
Predicted 0.472
Node 4
Mean 0.000
Std. Dev. 0.000
n 37
% 2.0
Predicted 0.000
Attorney involvement - plaintiff
Adj. P-value=0.008, F=7.177,
df1=1, df2=685
Y
N
Node 5
Mean 0.224
Std. Dev. 0.417
n 662
% 36.4
Predicted 0.224
Node 6
Mean 0.000
Std. Dev. 0.000
n 25
% 1.4
Predicted 0.000

CART

- Classification and regression trees
- Does binary (two-way) splits
- For numeric dependent variables, splits on variable creating greatest reduction in sum of squared errors
- For categorical variable, usually Gini index or entropy
- Keeps splitting until no significant reduction in sum squared errors

CART Tree for Paid Loss Severity



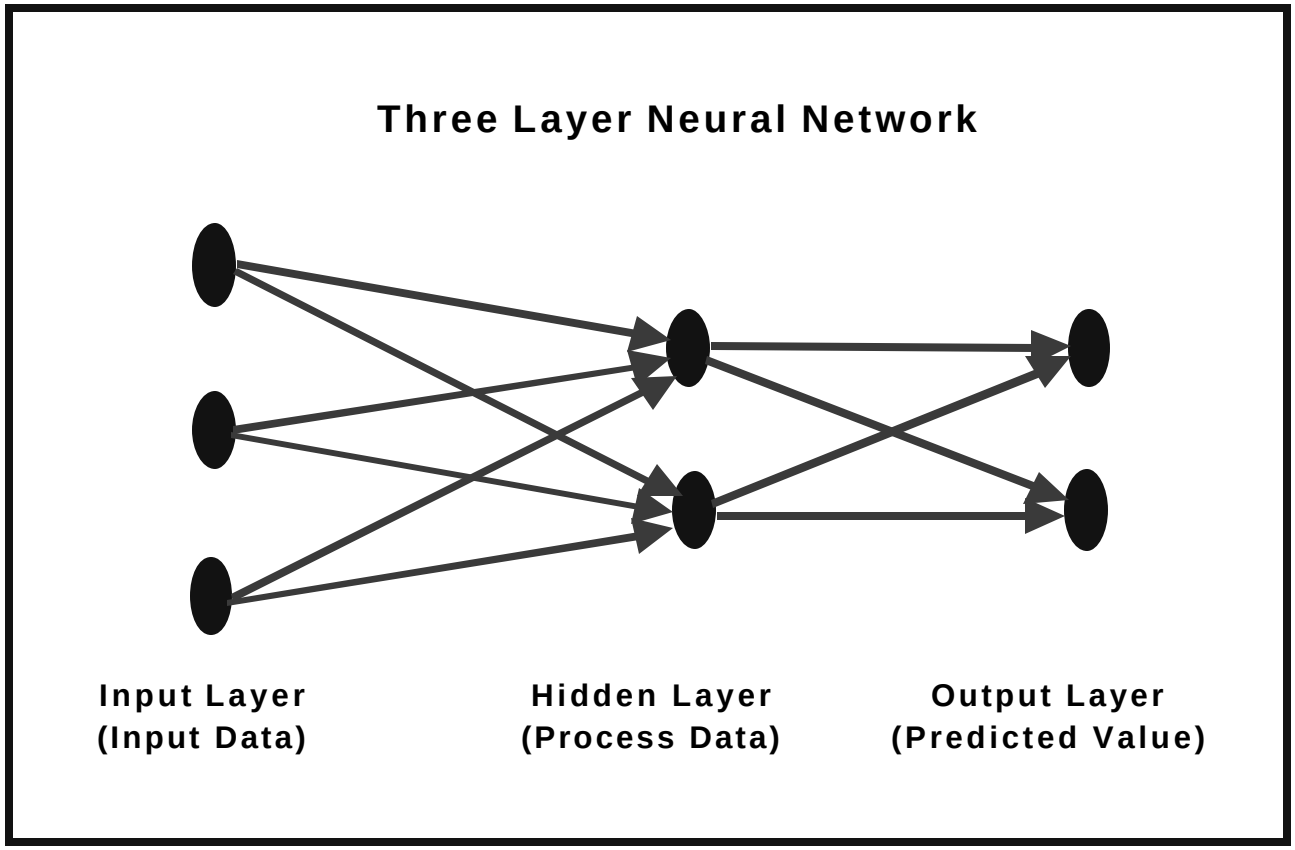
Neural networks

- Developed in artificial intelligence discipline
- Based on how neurons function in the brain
- Can be viewed as a generalization of regression

Neural Networks

- Also minimizes squared deviation between fitted and actual values
- Can be viewed as a non-parametric, non-linear regression

The Feedforward Neural Network



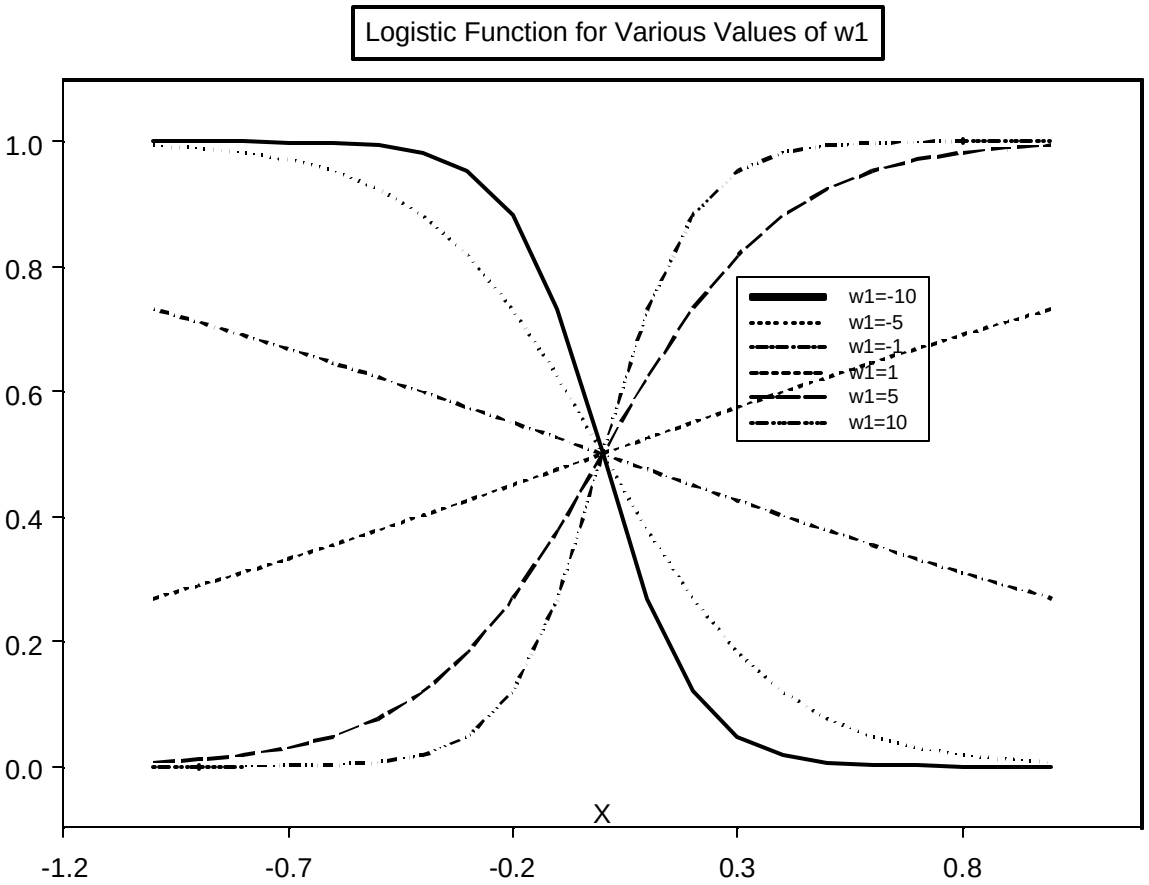
The Activation Function

- The sigmoid logistic function

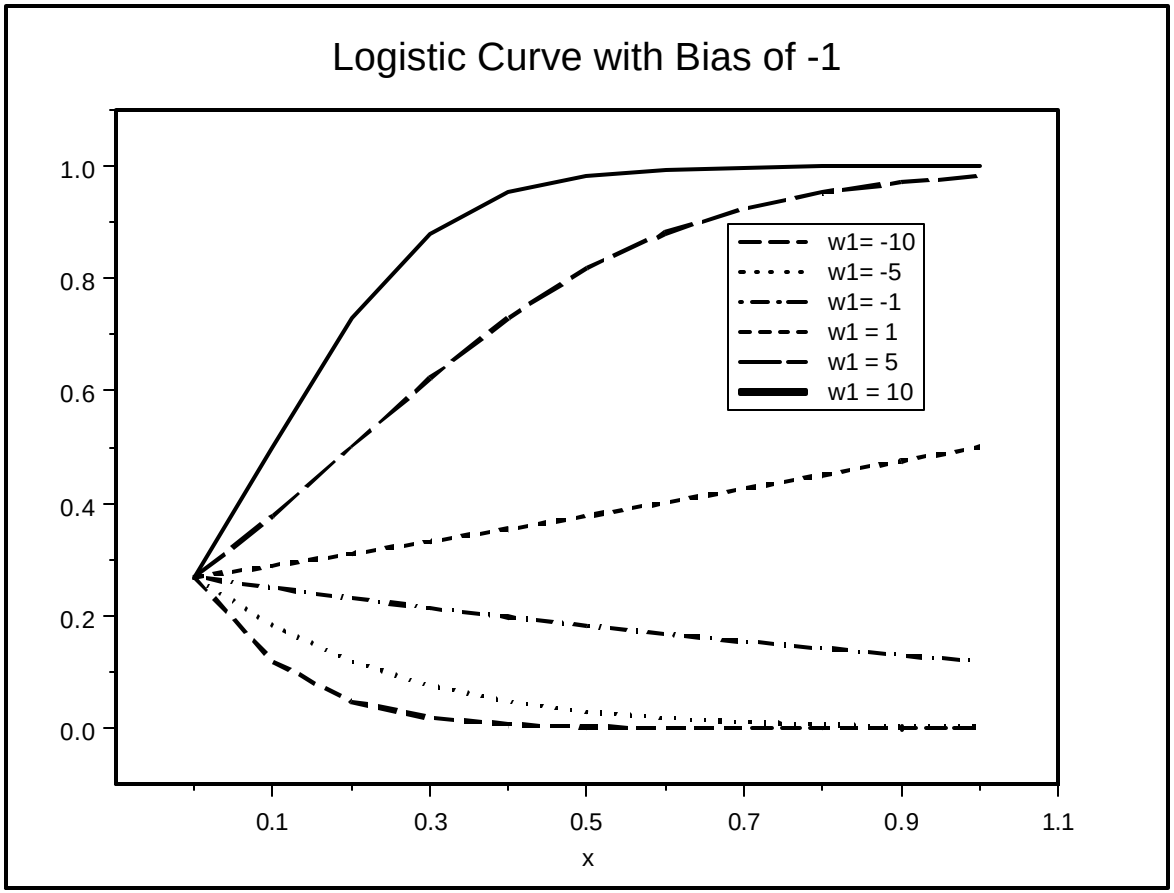
$$f(Y) = \frac{1}{1 + e^{-Y}}$$

$$Y = w_0 + w_1 * X_1 + w_2 X_2 \ldots + w_n X_n$$

The Logistic Function for Nonlinear Modeling



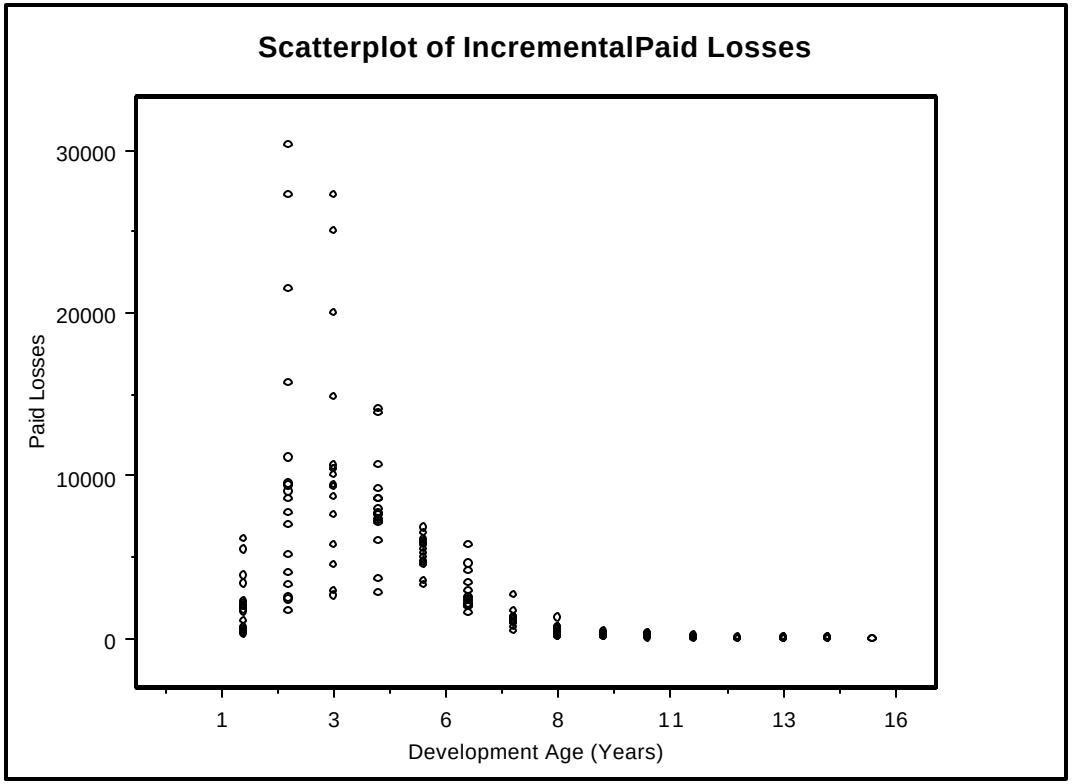
Variety of Shapes with Logistic Curve



Neural Network Software

- There are some inexpensive Excel add-ins
  - Good for becoming acquainted with method
  - Neural XL
  - Brainmaker
- R (www.r-project.org) free statistical language
  - Load nnet or neural library

Development Example: Incremental Payments Used for Fitting



Two Methods for Fitting Development Curve

- Neural Networks
  - Simpler model using only development age for prediction
  - More complex model using development age and accident year
- GLM model
  - Example uses Poisson regression
  - Like OLS regression, but does not require normality
  - Fits some nonlinear relationships
  - See England and Verrall, PCAS 2001

The Chain Ladder Model

Cumulative paid:

$$D_{ij} = \sum_{k=1}^{j} C_{ik}$$

Age to age factor:

$$l_{ij} = \frac{D_{i,j+1}}{D_{ij}}$$

Estimate of age to age factor using mean:

$$l_j = \frac{\sum_{i=1}^{n} l_{ij}}{n}$$

Common Approach: The Deterministic Chain Ladder

Estimate of paid at 24 months:

$$C_{24} = D_{12}\, l_{12} - D_{12}$$

Estimate of Ultimate Paid:

$$D_{iu} = D_{ij} \prod_{k=j}^{u} l_{ik}$$

GLM Model

A Stochastic Chain Ladder Model

Poisson Model:

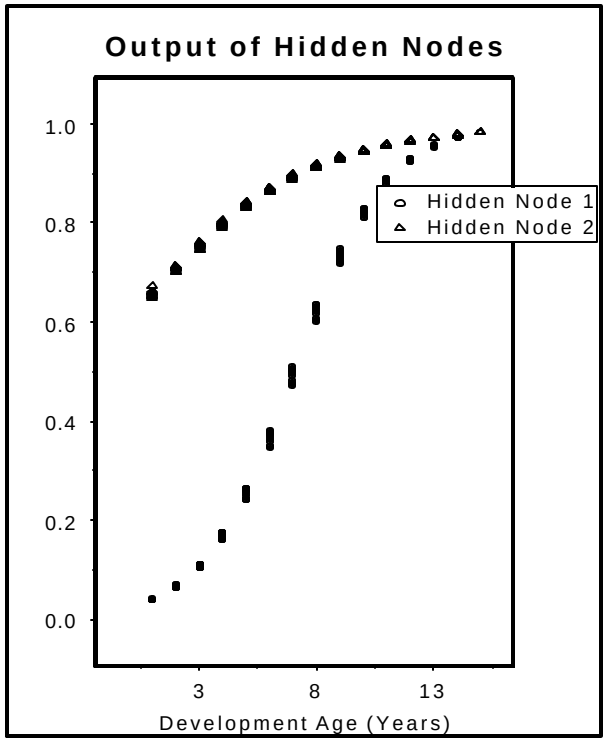
$$E(C_{ij}) = m_{ij} = x_i y_j$$

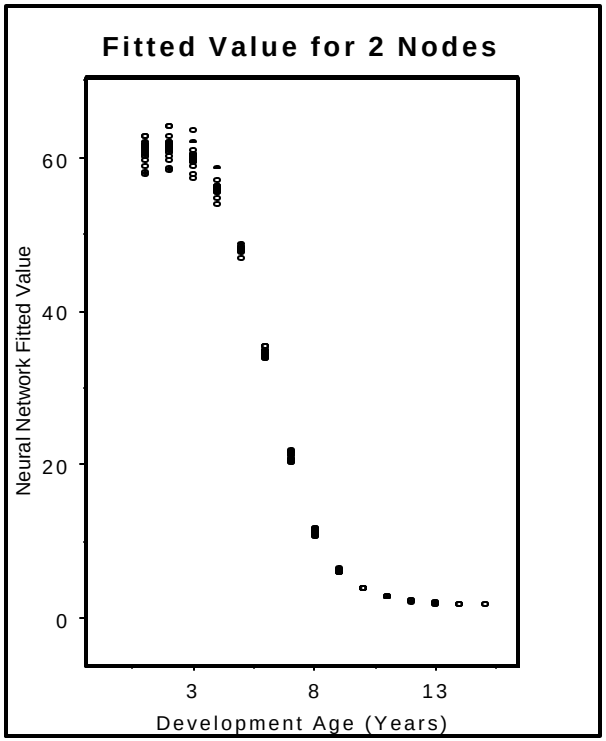
$$Var[C_{ij}] = f x_i y_j$$

$$\sum_{k=1}^{n} y_k = 1$$

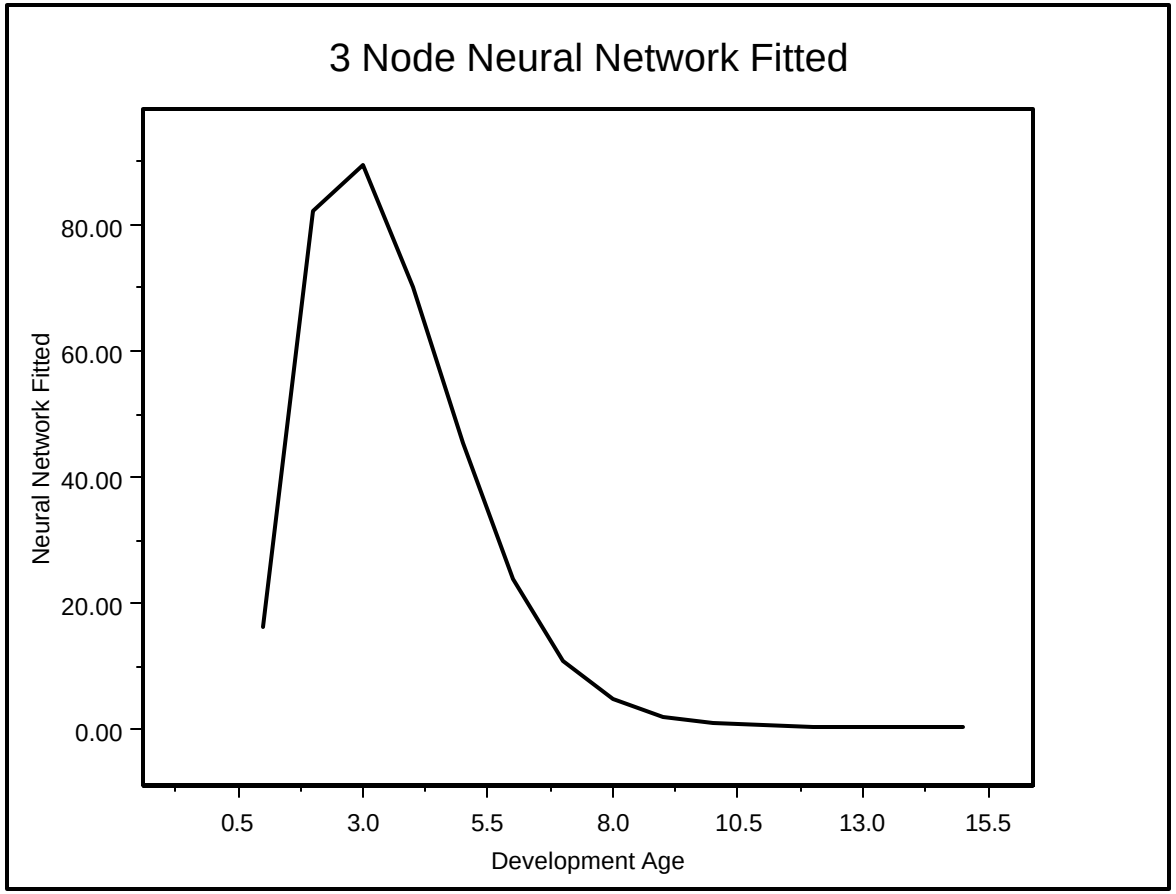
Data often normalized by dividing by an exposure base

Hidden Nodes for Paid Chain Ladder Example





NN Chain Ladder Model with 3 Nodes

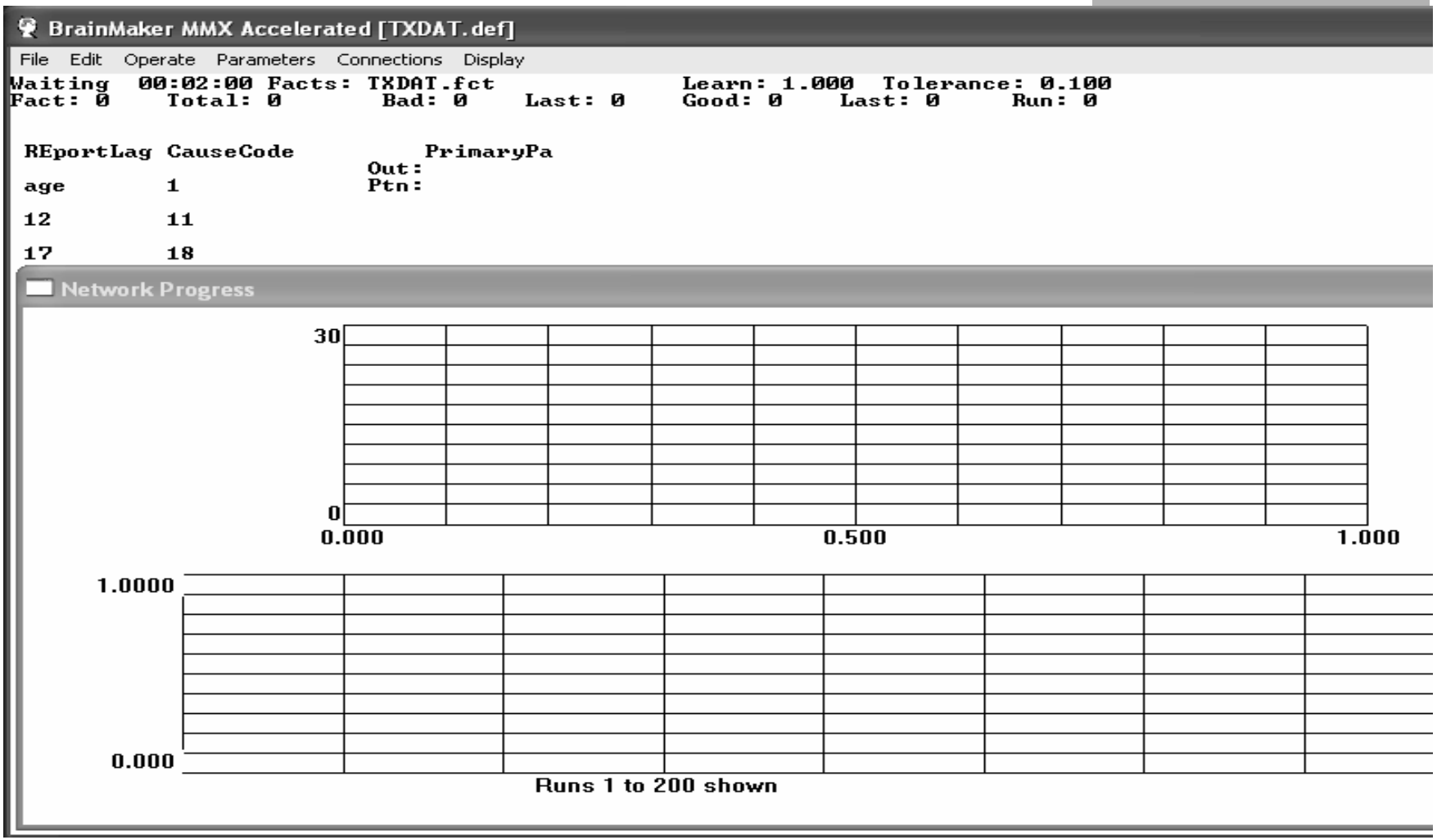


Universal Function Approximator

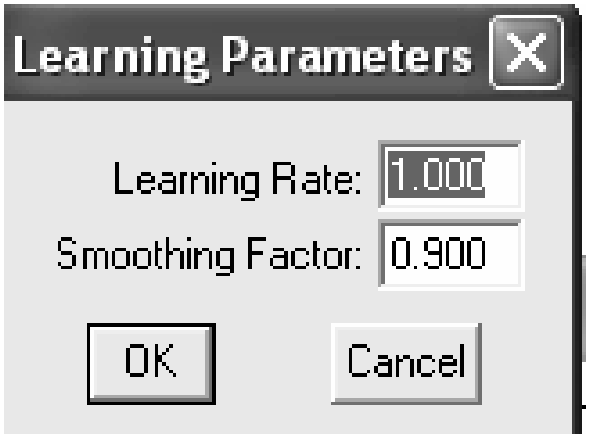
- The feedforward neural network with one hidden layer is a universal function approximator

- Theoretically, with a sufficient number of nodes in the hidden layer, any continuous nonlinear function can be approximated

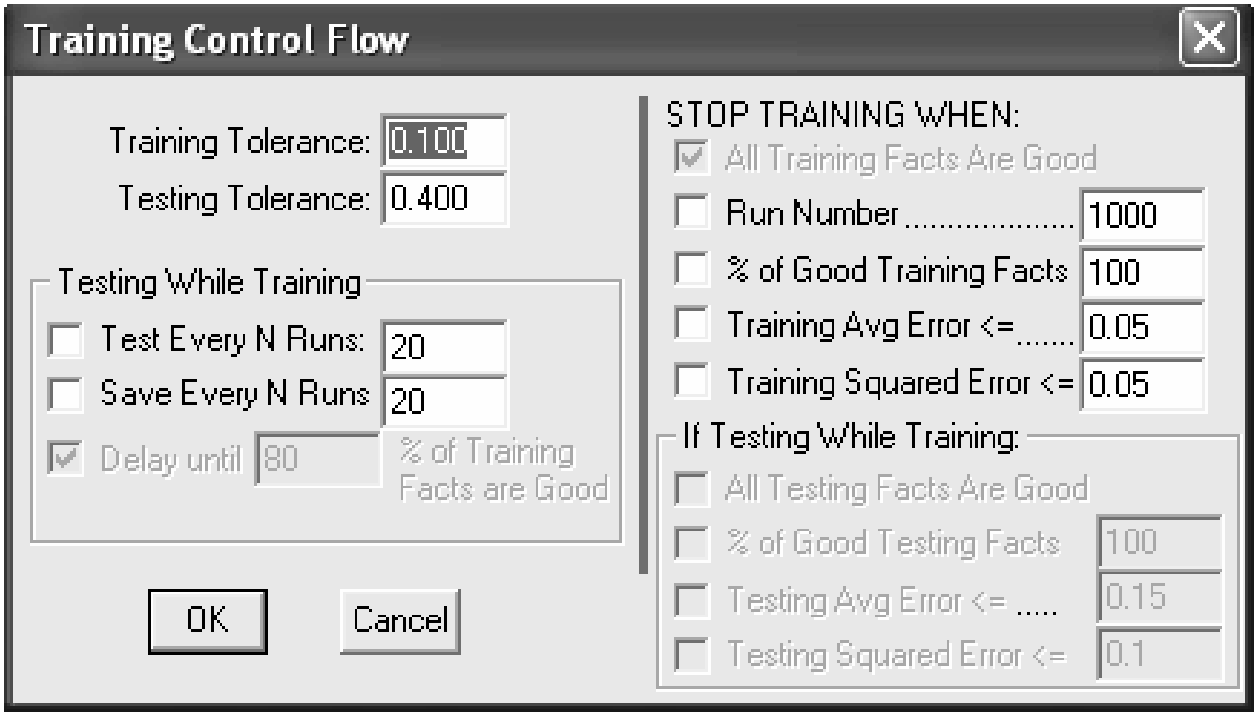
Excel Add-In Neural Network: Set-up



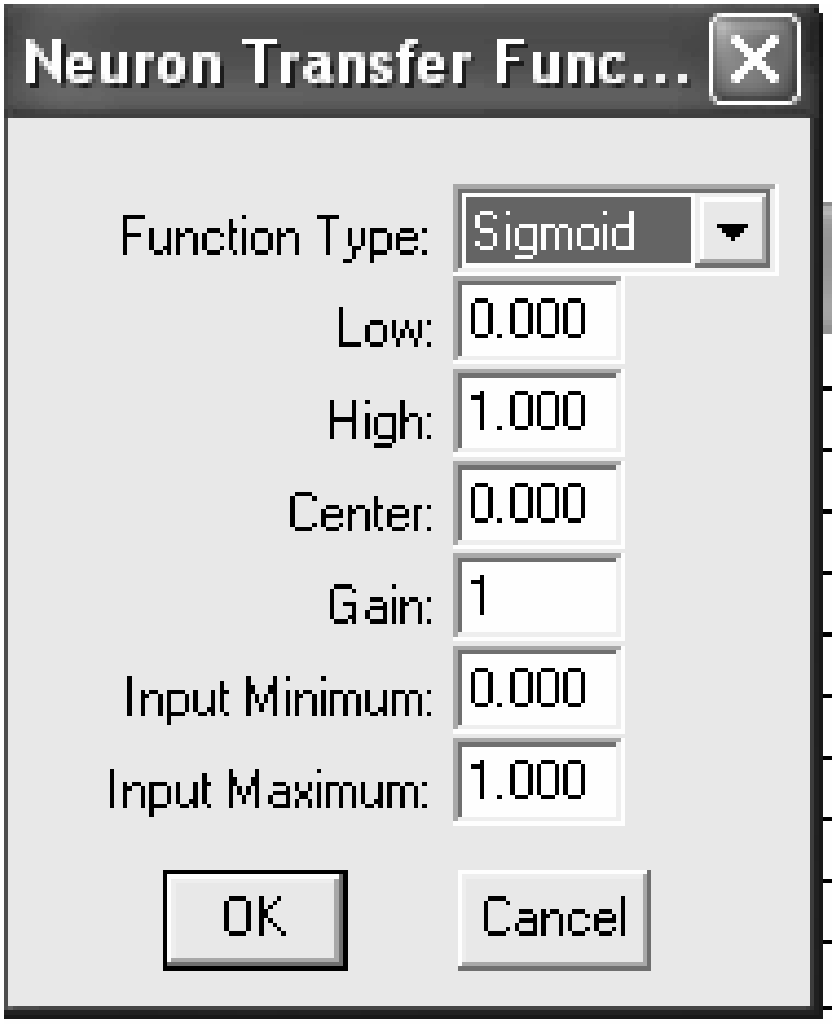
Learning Rate



Training Parameters

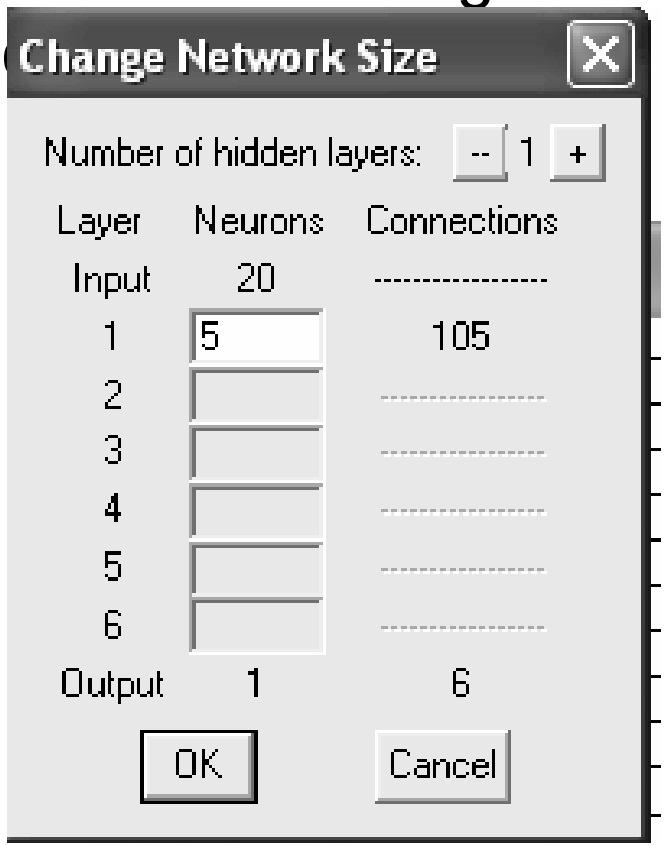


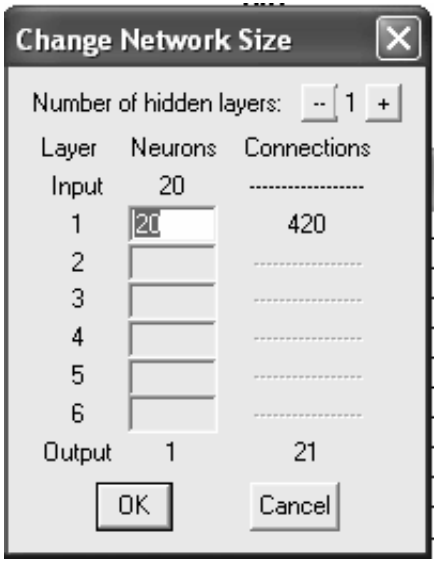
Activation Function and Scaling



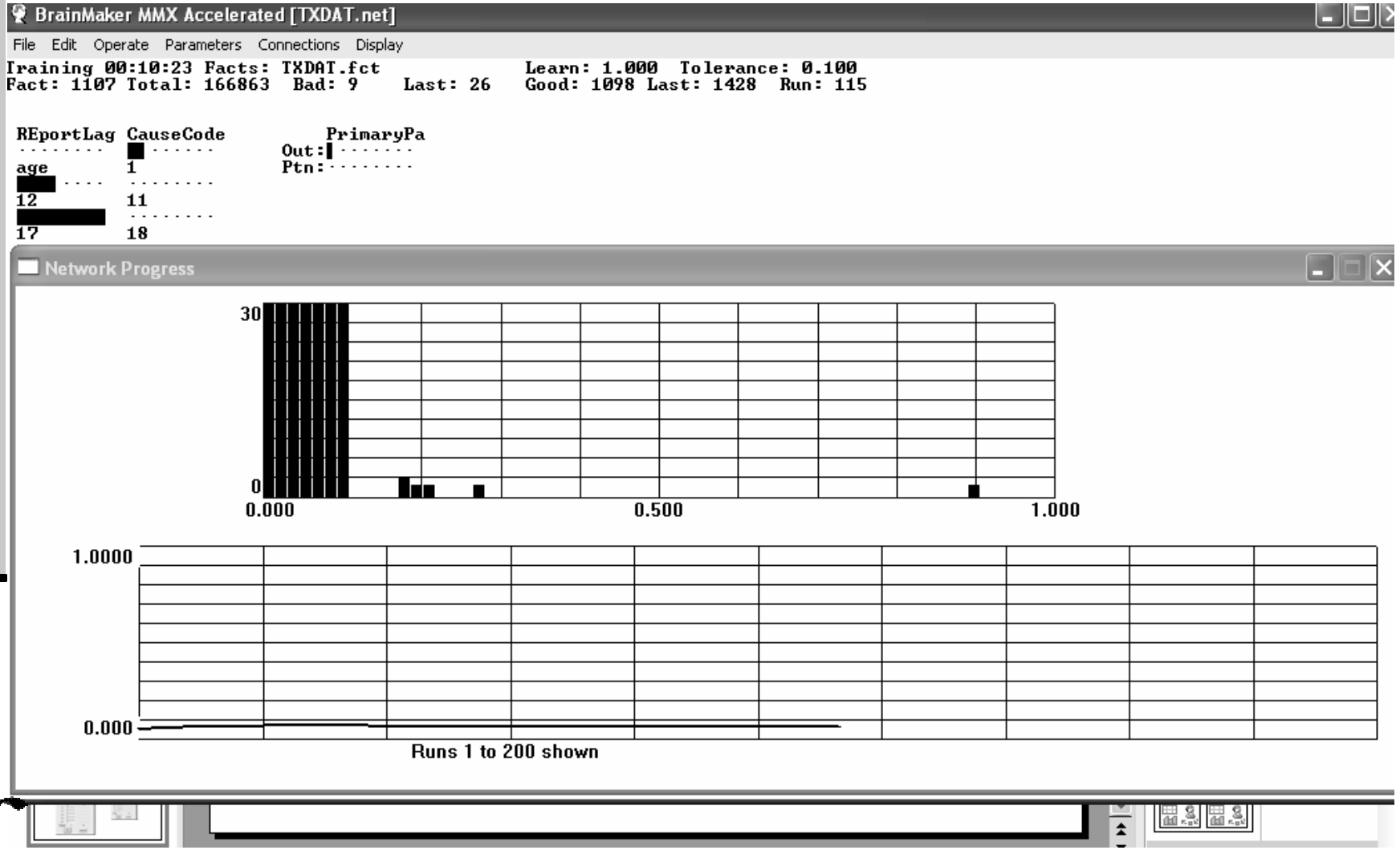
How many Hidden Neurons?

- Too many – overparameterize
- Too few – don't get g

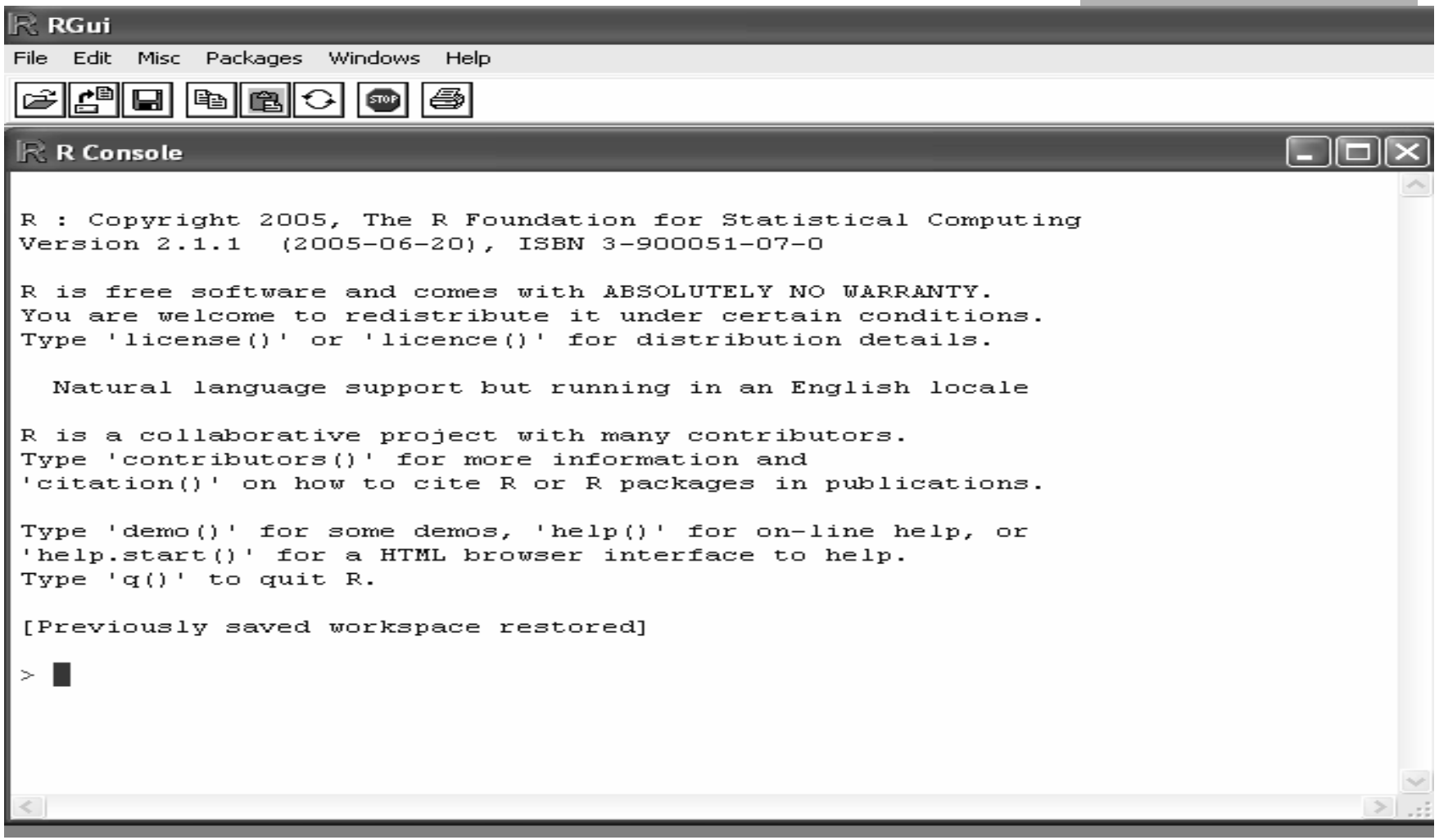




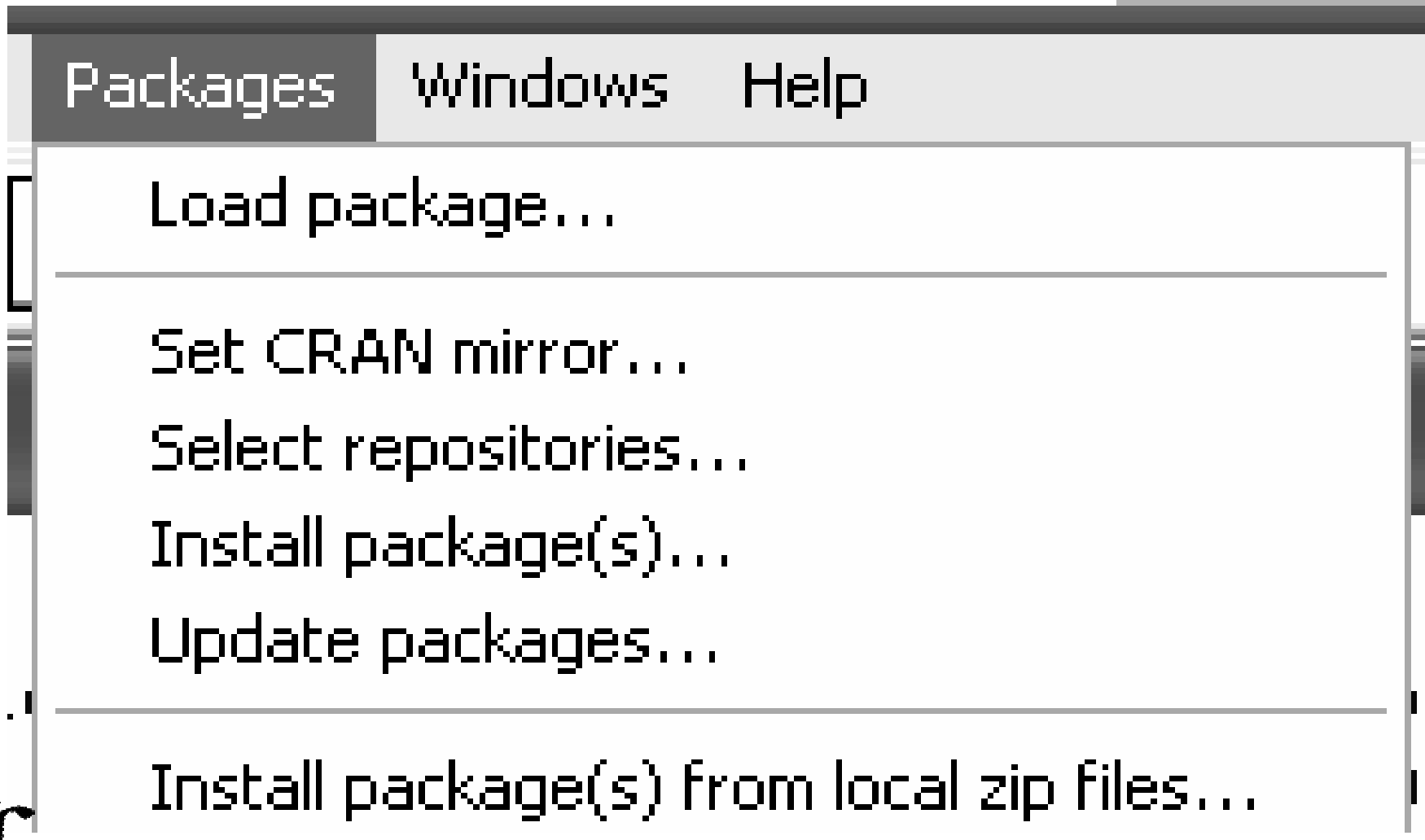
Training Network



Alternative Load R



Load nnet library



R Code for Neural Networks

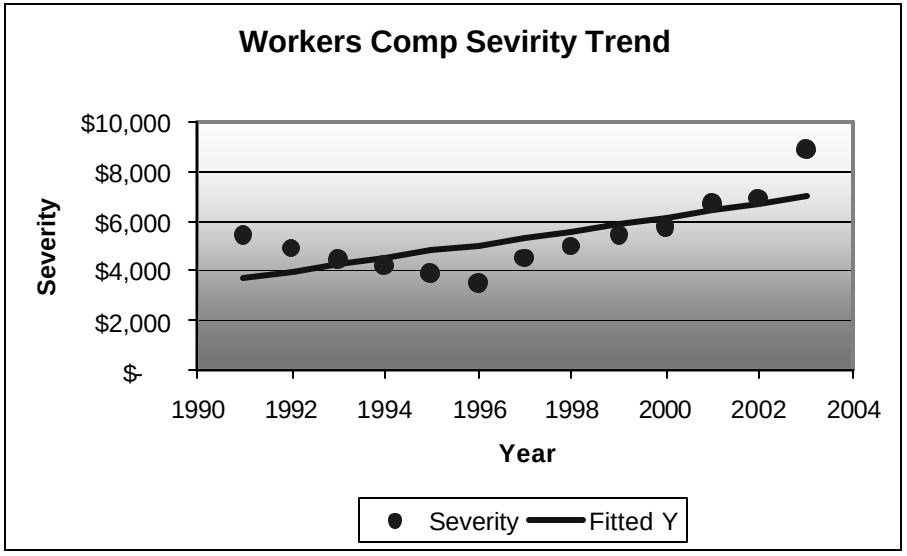
- Txdata<-
  read.table('C:/Seminar/TxDat.dat',header=T)
- Paid.nnet<-
  nnet(PrimaryPaid~REportLag+age+Injury+Cause,data=Txdata,size=5,linout=T)

Fitting R Neural Network

```
> Paid.nnet<-nnet(PrimaryPaid~REportLag+age+Injury+Cause,data=Txdata)
Error in nnet.default(x, y, w, ...) : argument "size" is missing, with no defau$
> ?nnet
> Paid.nnet<-nnet(PrimaryPaid~REportLag+age+Injury+Cause,data=Txdata,size=5,lin$
# weights:  171
initial  value 289704047667119.750000
final  value 212668556210510.690000
converged
>
```

Review from Last Workshop: Regression for Prediction

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



Test the model

- Hold out part of the sample, say one third and test its performance after model has been fit
- Cross-validation

Which Model to Use?

- Classification
  - Naïve bayes makes simplifying assumption of independence
  - Logistic regression is one of most frequently used. Its use requires special software (such as R)
  - To capture data complexities such as nonlinearities, trees are easy to understand, common method
  - To capture data complexities neural networks also common method, but it is harder to understand

Which model to use?

- Numeric variables
  - Trees usually easier to understand and explain
  - Neural networks often give better fit
  - More recently developed methods (ensemble models) seem to outperform neural networks

Data Mining Library Recommendations

- Berry, Michael J. A., and Linoff, Gordon, *Data Mining Techniques*, John Wiley and Sons, 1997
- Hosmer D and Lemshow, *Applied Logistic Regression,* Wiley
- Francis, Louise, 2006 Data Quality/Management Call Paper Program, “Distinguishing the Forest from the Trees”, www,caact.org
- Type naïve bayes into search at www.wikipedia.org
- Francis, Louise, 2001, “Neural Networks Demystified” at www.casact.org/aboutcas/mdiprize.htm