Introduction to Generalized Linear Models

2006 CAS Predictive Modeling Seminar

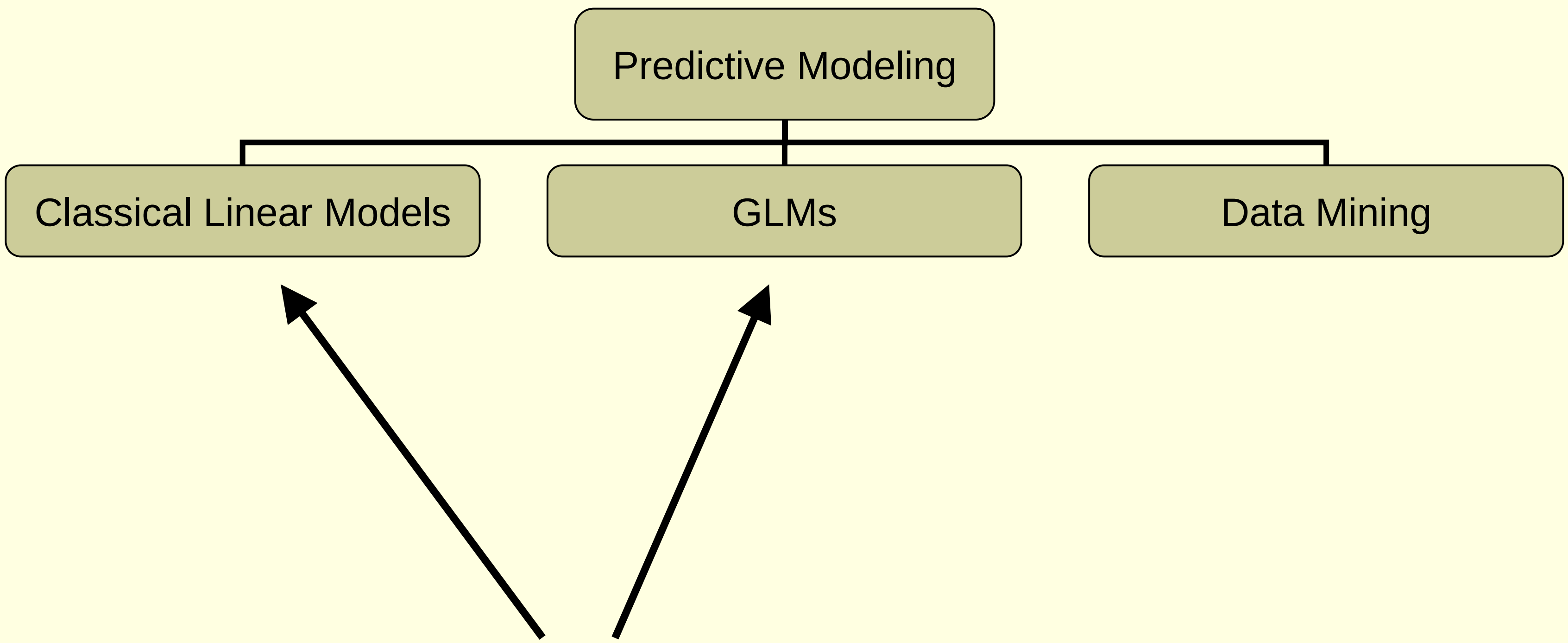
Prepared by
Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com
Louise_francis@msn.com

October 4, 2006

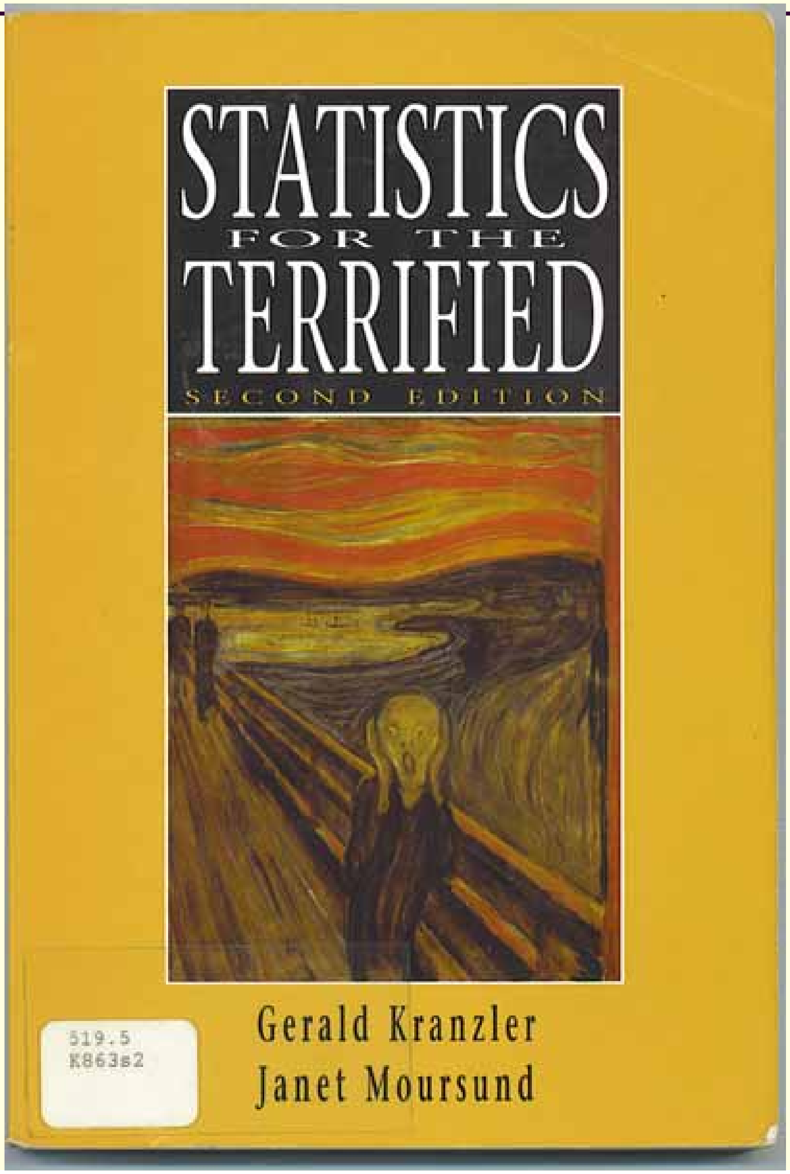
Objectives

- Gentle introduction to **Linear Models**
- Illustrate some simple applications of linear models
- Address some practical modeling issues
- Show features common to LMs and GLMs

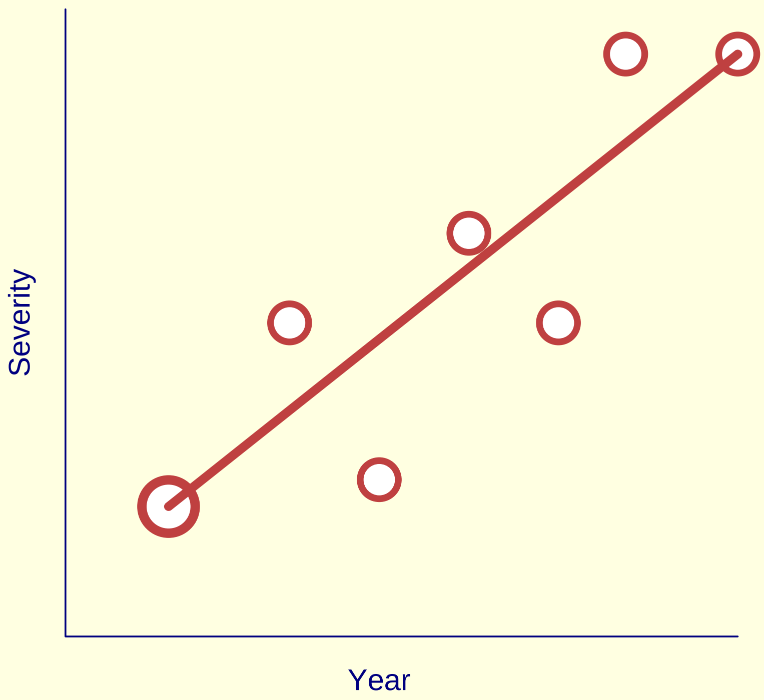
Predictive Modeling Family



Many Aspects of Linear Models are Intuitive

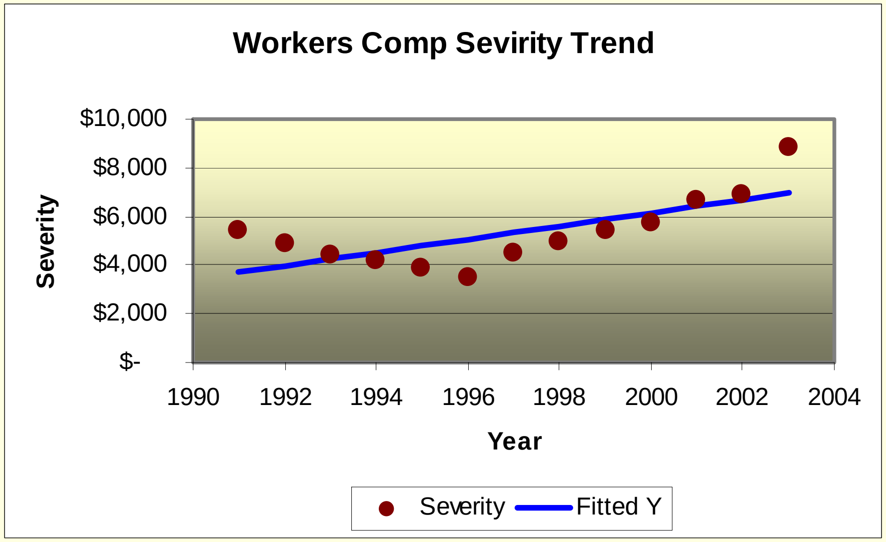


An Introduction to Linear Regression



Intro to Regression Cont.

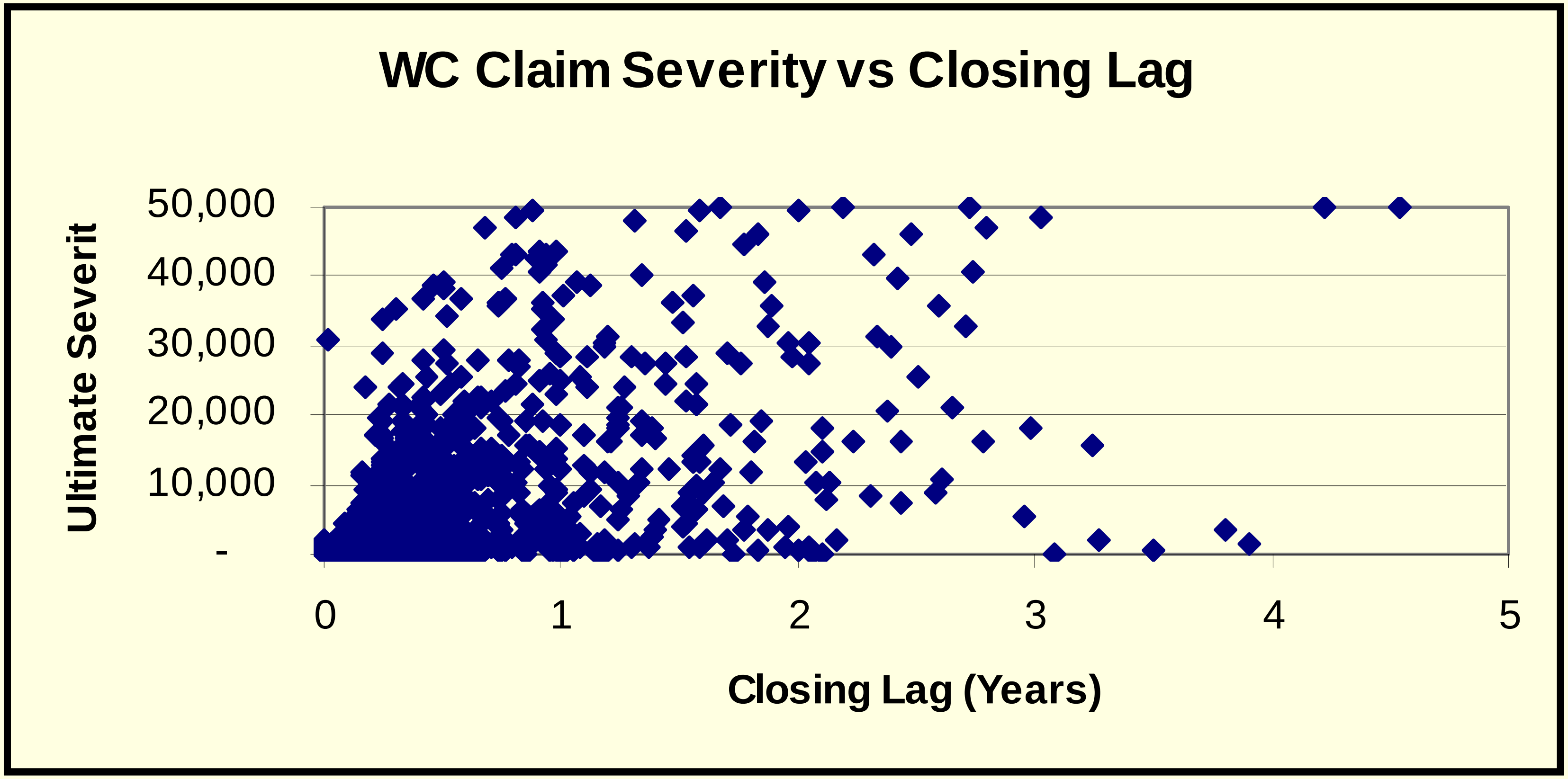
- Fits line that minimizes squared deviation between actual and fitted values
- $$\min\left(\sum (Y_i - \widehat{Y})^2\right)$$



Some Workers Compensation Data

- Ultimate Severity
- Lags
  - Closing
  - Report
- Claim Type
  - Med Only
  - Fast Track
  - Lost Time
- Injury
  - Sprain, strain, cut, etc.

Simple Illustration
Severity vs. Closing Lag



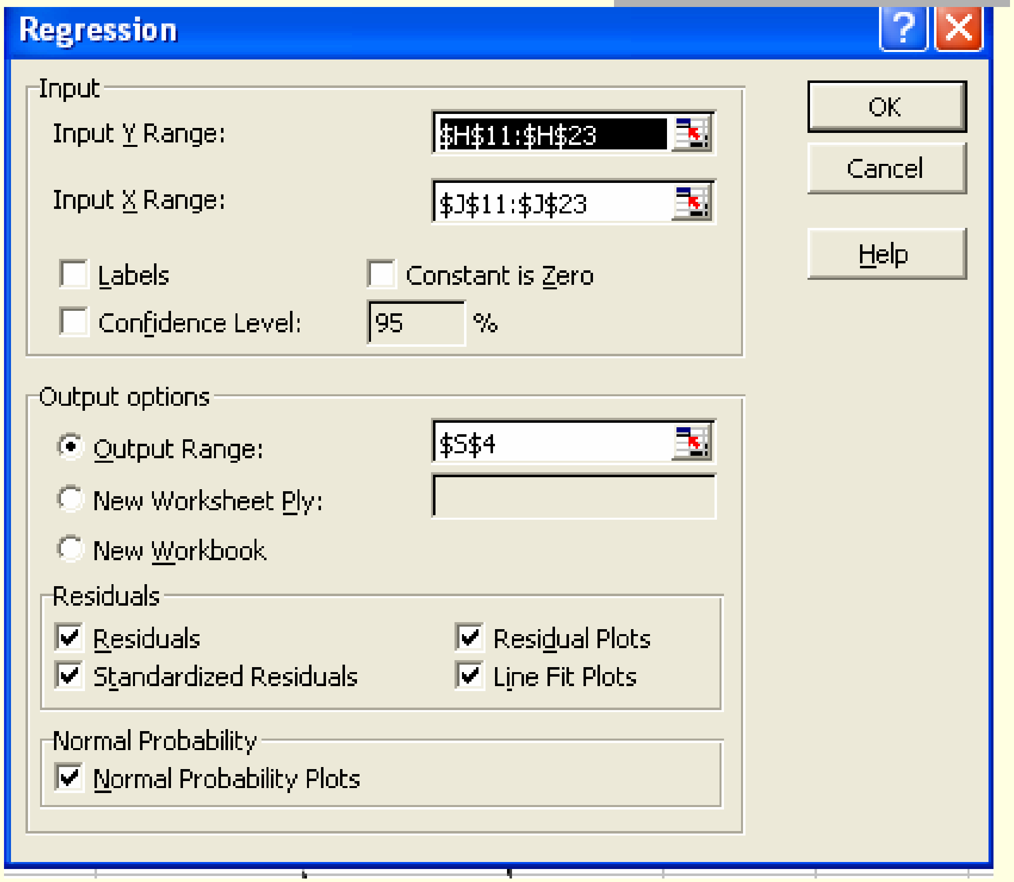
How Strong Is Linear Relationship?: Correlation Coefficient

- Varies between -1 and 1
- Zero = no linear correlation

	Severity	*Report Lag*	*Closing Lag*
Severity	1.000		
Report Lag	(0.019)	1.000	
Closing Lag	0.645	0.000	1.000

Excel Does Regression

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



How Good is the fit?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6351
R Square	0.4034
Adjusted R Square	0.4033
Standard Error	13307
Observations	5631

First Step: Compute residual

- Residual = actual – fitted

Actual Severity	*Predicted Severity*	*Residuals*
-	(2,965)	2,965
272	444	(173)
752	368	383
762	444	318

- Sum the square of the residuals (SSE)
- Compute total variance of data with no model (SST)

Goodness of Fit Statistics

- R^2: (SSE Regression/SS Total)
  - percentage of variance explained
- Adjusted R^2
  - R^2 adjusted for number of coefficients in model

    - Note SSE = Sum squared errors
    - MS id Mean Square Error

R^2 Statistic

Regression Statistics	
Multiple R	0.6351
R Square	0.4034
Adjusted R Square	0.4033
Standard Error	13,307
Observations	5,631

Significance of Regression

- F statistic:
  - (Mean square error of Regression/Mean Square Error of Residual)

ANOVA (Analysis of Variance) Table

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7,036	7,036	1,404	0
Residual	5,629	28,211	5		
Total	5,630	35,247			

Goodness of Fit Statistics

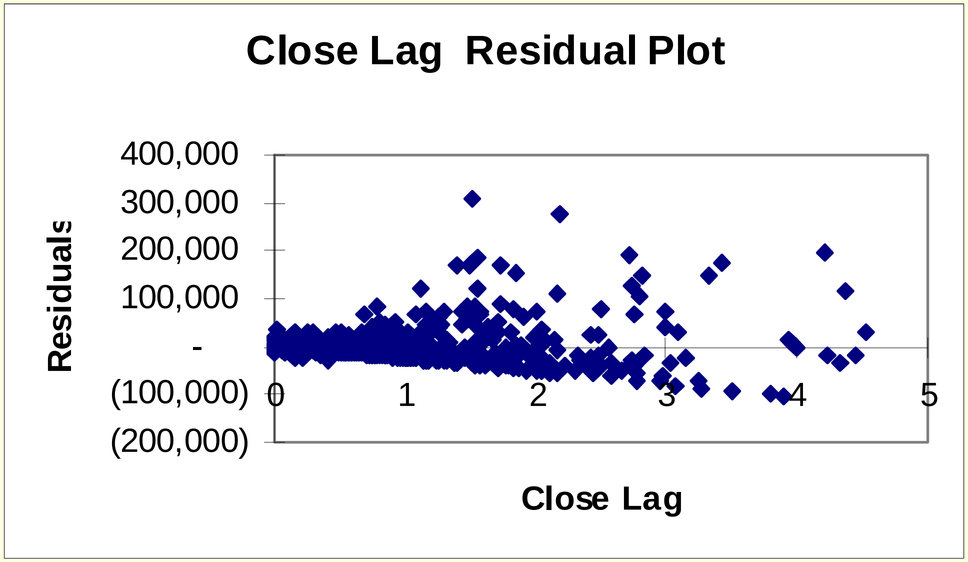
- T statistics: Uses SE of coefficient to determine if it is significant
  - SE of coefficient is a function of ***s*** (mean square error of regression)
  - Uses T-distribution for test
  - It is customary to drop variable if coefficient not significant

T-Statistic: Are the Intercept and Coefficient Significant?

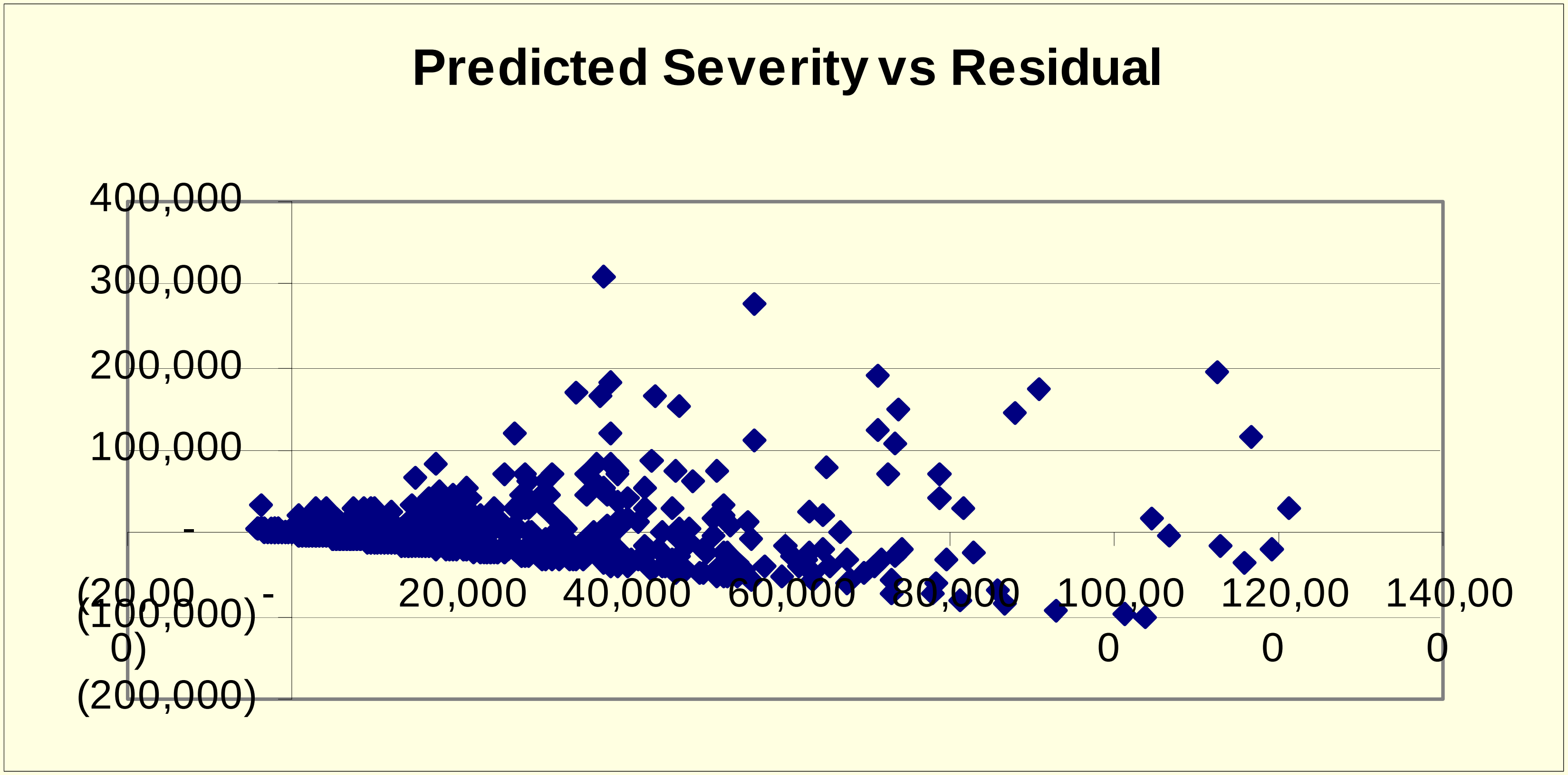
Parameter	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	(4,025.1)	218.1	(18.5)	0.0	(4,452.7)	(3,597.5)
Closing Lag	27,667.9	448.5	61.7	-	26,788.6	28,547.1

Other Diagnostics: Residual Plot Independent Variable vs. Residual

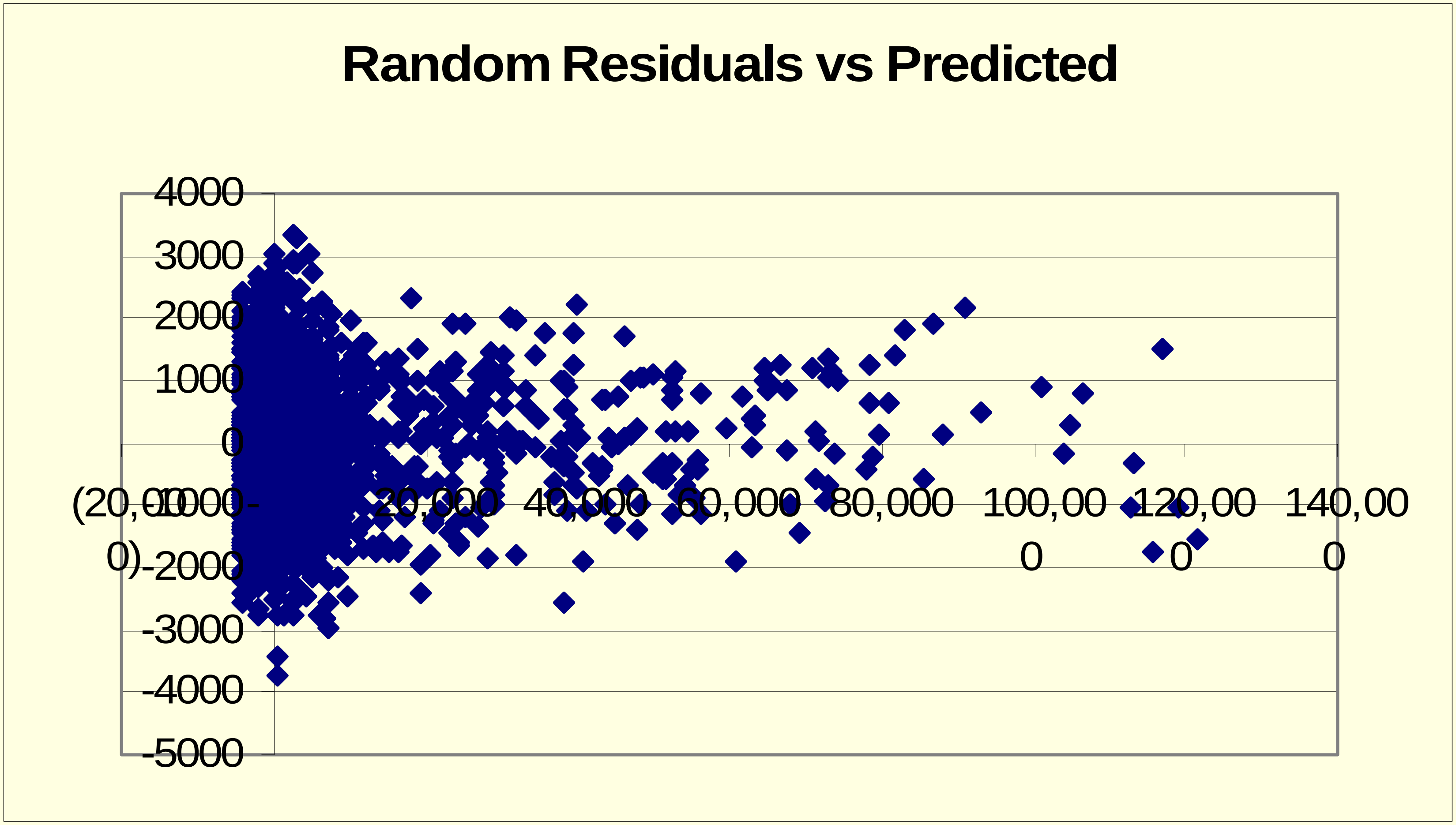
- Points should scatter randomly around zero
- If not, a straight line probably is not be appropriate



Predicted vs. Residual



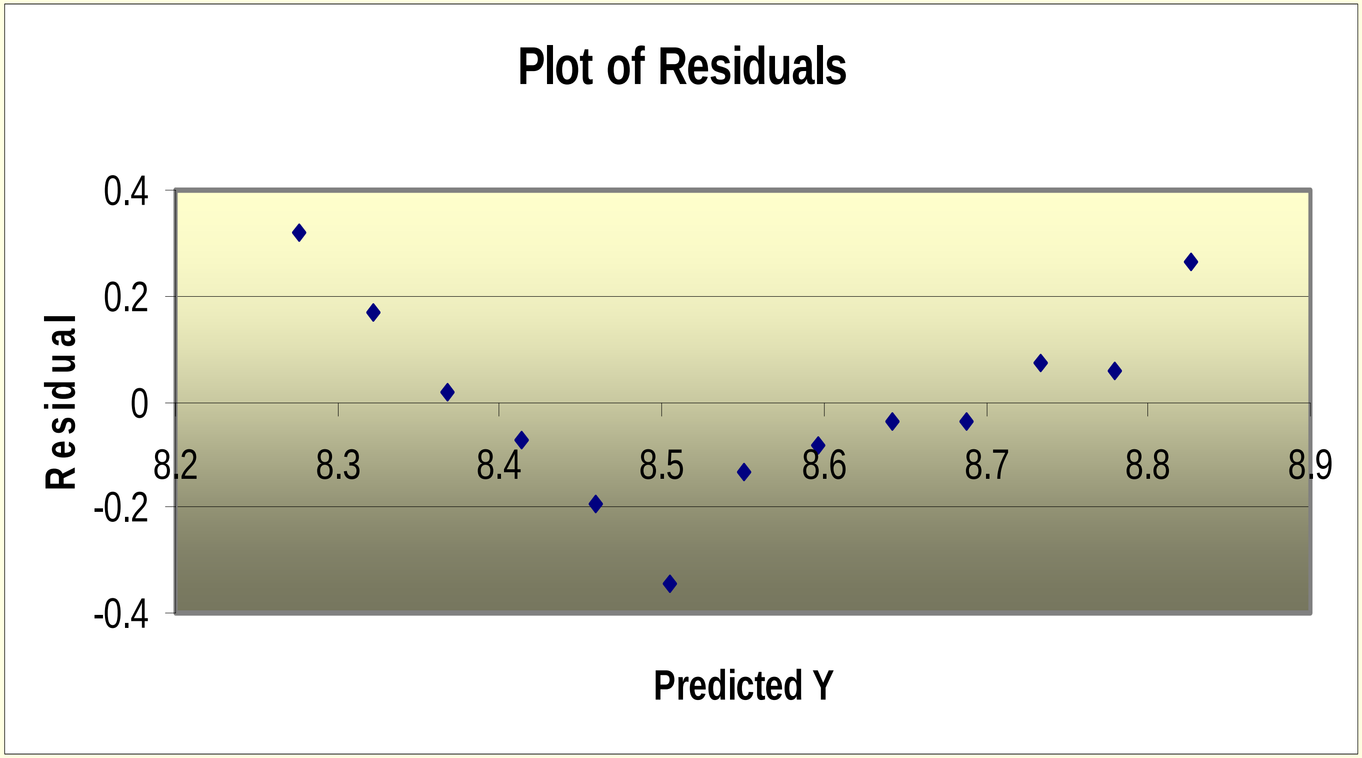
Random Residual



What May Residuals Indicate?

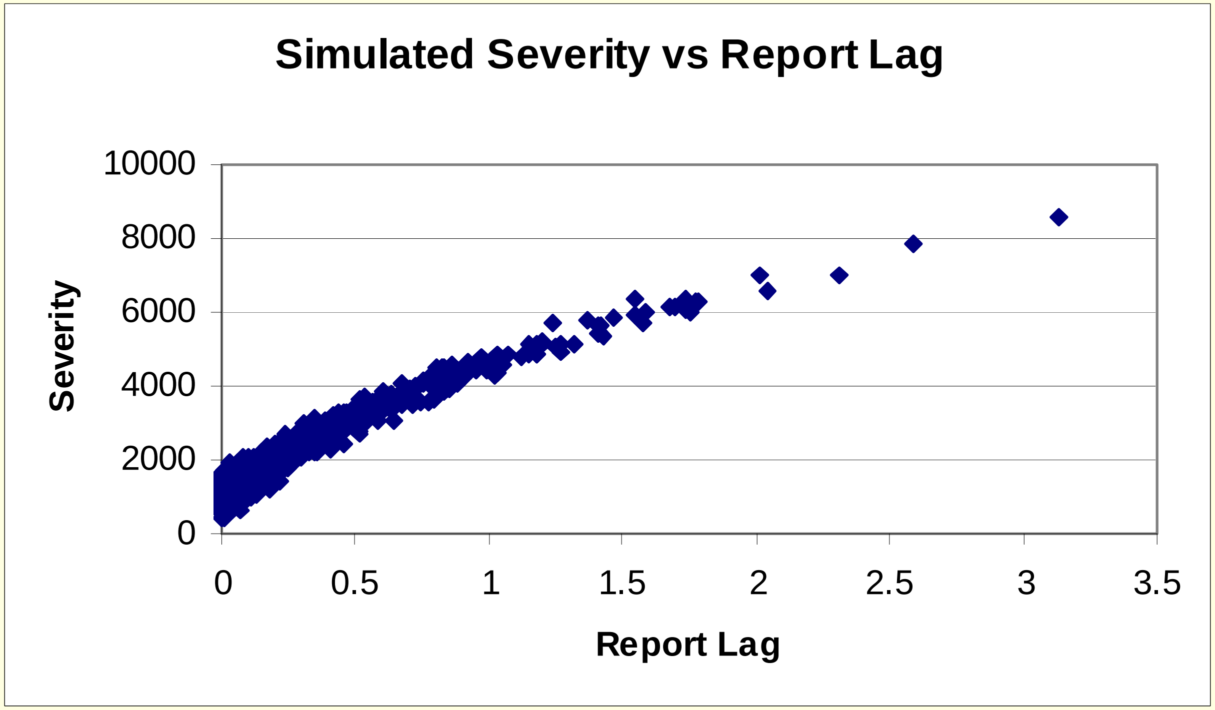
- If absolute size of residuals increases as predicted increases, may indicate non-constant variance
  - may indicate need to log dependent variable
  - Use weighted regression
    - Weight inversely proportional to variance
- May indicate a nonlinear relationship

Non-Linear Relationship



Non-Linear Relationships

- Suppose Relationship between dependent and independent variable is non-linear?
- Linear regression requires a linear relationship



Transformation of Variables

- Apply a transformation to either the dependent variable, the independent variable or both
- Examples:
  - $Y' = \log(Y)$
  - $X' = \log(X)$
  - $X' = 1/X$
  - $Y'=Y^{1/2}$

Transformation of Variables

- Suppose Severity is a function of the log of report lag
  - Compute X' = log(Report Lag)
  - Regress Severity on X'

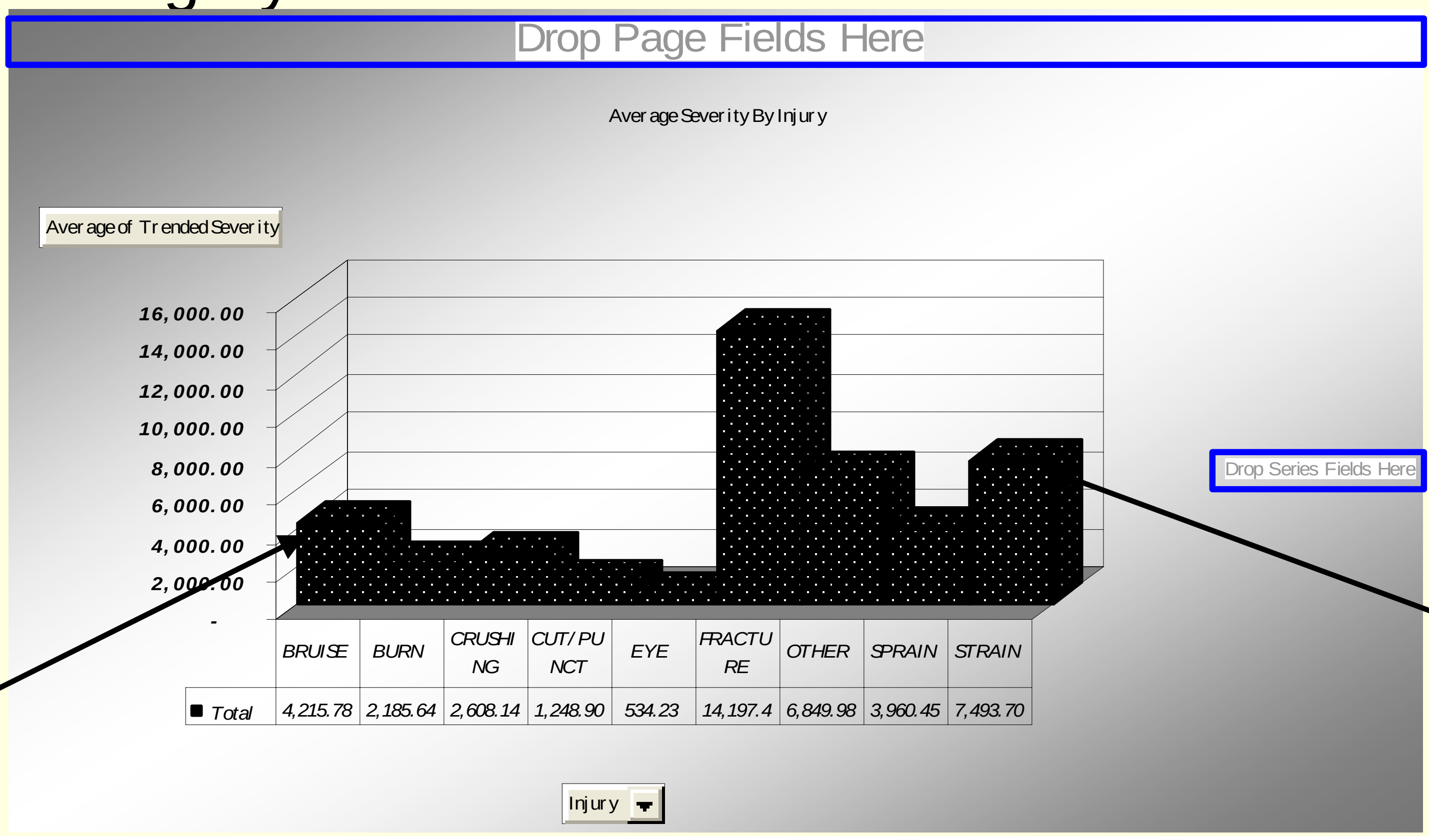
	Coefficients	*Standard Error*	*t Stat*
Intercept	1003.58	5.01	200.43
Log Report Lag	12049.13	78.01	154.46

Categorical Independent Variables: The Other Linear Model: ANOVA

Average of Trended Severity	
Injury	Total
BRUISE	4,215.78
BURN	2,185.64
CRUSHING	2,608.14
CUT/PUNCT	1,248.90
EYE	534.23
FRACTURE	14,197.49
OTHER	6,849.98
SPRAIN	3,960.45
STRAIN	7,493.70
Grand Total	4,650.76

Model

- Model is Model $Y = a_i$, where i is a category of the independent variable. a_i is the mean of category i.



$Y=a_1$

$Y=a_9$

Two Categories

- Model Y = a_i, where i is a category of the independent variable
- In traditional statistics we compare a_1 to a_2

	Data	
SPRAIN/STRAIN	Average of Trended Severity	Count of Trended Severity
OTHER	3,793	3,086
SPRAIN/STRAIN	6,869	1,193
Grand Total	4,651	4,279

If Only Two Categories: T-Test for test of Significance of Independent Variable

	Variable 1	*Variable 2*
Mean	3,793	6,869
Variance	270,835,811	672,171,797
Observations	3,086	1,193
Pooled Variance	382,688,160	
Hypothesized Mean D	-	
df	4,277	
t Stat	(4.61)	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.00	
t Critical two-tail	1.96	

More Than Two Categories

- Use F-Test instead of T-Test
- With More than 2 categories, we refer to it as an Analysis of Variance (ANOVA)

Fitting ANOVA With Two Categories Using A Regression

- Create A Dummy Variable for Sprain/Strain
- Variable is 1 of SPRAIN/STRAIN, and 0 Otherwise

Severity	SPRAIN/STRAIN	Dummy Variable
-	OTHER	0
271.53	OTHER	0
751.71	SPRAIN/STRAIN	1
762.08	OTHER	0
796.75	OTHER	0

More Than 2 Categories

- If there are k Categories:
- Create k-1 Dummy Variables
  - Dummy_i = 1 if claim is in category i, and is 0 otherwise
- The k^{th} Variable is 0 for all the Dummies
- Its value is the intercept of the regression

Design Matrix

Severity	Injury	Dummy 1	Dummy 2	Dummy 3	Dummy 4	Dummy 5	Dummy 6	Dummy 7	Dummy 8
-	BRUISE	0	1	0	0	0	0	0	0
271.53	OTHER	0	0	0	0	0	0	0	0
751.71	STRAIN	0	0	1	0	0	0	0	0
762.08	FRACTURE	0	0	0	0	1	0	0	0
796.75	CUT/PUNCT	1	0	0	0	0	0	0	0
382.20	BRUISE	0	1	0	0	0	0	0	0
171.35	EYE	0	0	0	0	0	0	1	0

Regression Output for Categorical Independent

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.16
R Square	0.03
Adjusted R Square	0.02
Standard Error	19,621.92
Observations	4,112.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	4.36E+10	5.45E+09	14	0
Residual	4103	1.58E+12	3.85E+08		
Total	4111	1.62E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6,410.86	954.05	6.72	0.00	4,540.40	8,281.32
Dummy 1	(5,130.72)	1,130.93	(4.54)	0.00	(7,347.96)	(2,913.48)
Dummy 2	(2,153.48)	1,147.89	(1.88)	0.06	(4,403.96)	97.00
Dummy 3	1,140.73	1,148.45	0.99	0.32	(1,110.86)	3,392.31
Dummy 4	(2,332.76)	1,683.84	(1.39)	0.17	(5,634.00)	968.48
Dummy 5	8,148.78	1,716.79	4.75	0.00	4,782.94	11,514.61
Dummy 6	(4,205.91)	1,656.39	(2.54)	0.01	(7,453.34)	(958.48)
Dummy 7	(5,871.33)	2,299.01	(2.55)	0.01	(10,378.63)	(1,364.03)
Dummy 8	(5,532.85)	2,516.55	(2.20)	0.03	(10,466.65)	(599.04)

A More Complex Model Multiple Regression

- Let $Y = a + b_1*X_1 + b_2*X_2 + \ldots b_n*X_n+e$
- The X's can be numeric variables or categorical dummies

Multilple Regression

$Y = a + b1 * \text{Report lag} + c_i \text{Injury}_i + d_k \text{Claim Type}_k + e$

Regression Statistics	
Multiple R	0.39
R Square	0.15
Adjusted R Square	0.15
Standard Error	18,347.71
Observations	4,108.00

ANOVA

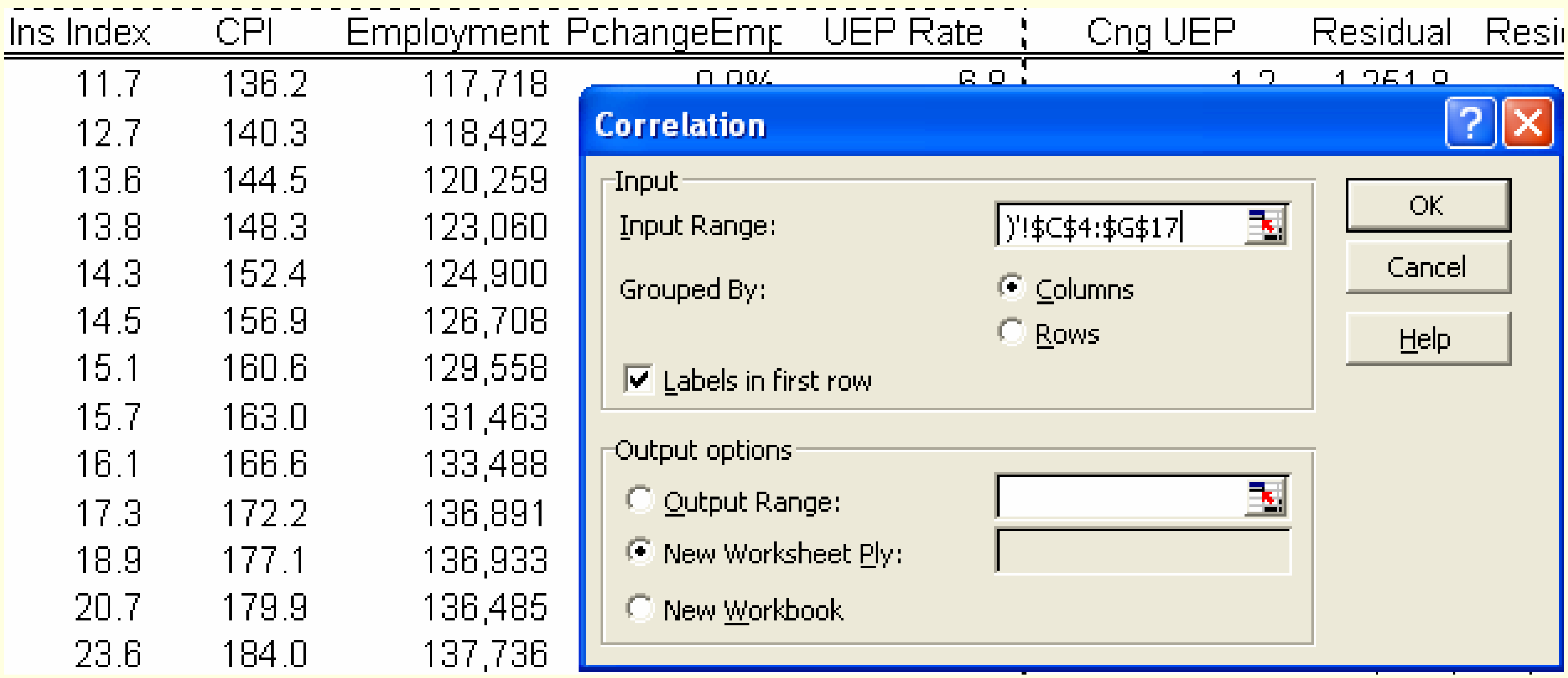
	df	*SS*	*MS*	*F*	*Significance F*
Regression	10.00	2.44066E+11	2.44E+10	72.50094	4.2148E-137
Residual	4,097.00	1.37921E+12	3.37E+08		
Total	4,107.00	1.62327E+12			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	18,831	1,039	18.12	0.00	16,793	20,869
Dummy 1	(1,132)	1,070	(1.06)	0.29	(3,230)	967
Dummy 2	(103)	1,078	(0.10)	0.92	(2,216)	2,009
Dummy 3	1,419	1,076	1.32	0.19	(689)	3,528
Dummy 4	(1,081)	1,578	(0.68)	0.49	(4,176)	2,013
Dummy 5	3,672	1,618	2.27	0.02	500	6,844
Dummy 6	(1,985)	1,553	(1.28)	0.20	(5,029)	1,059
Dummy 7	(1,023)	2,160	(0.47)	0.64	(5,258)	3,213
Dummy 8	(831)	2,362	(0.35)	0.72	(5,461)	3,799
Claim Type	(17,885)	734	(24.38)	0.00	(19,324)	(16,447)
Report Lag	134	2,228	0.06	0.95	(4,235)	4,502

More Than One Categorical Variable

- For each categorical variable
  - Create k-1 Dummy variables
  - K is the total number of categories
  - The category left out becomes the “base” category
  - It’s value is contained in the intercept
  - Model is $Y = a_i + b_j + \ldots + e$ or
  - $Y = u + a_i + b_j + \ldots + e$, where $a_i + b_j$ are offsets to u
    - e is random error term

Correlation of Predictor Variables: Multicollinearity



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

Similarities with GLMs

<u>Linear Models</u>

- Transformation of Variables
- Use dummy coding for categorical variables
- Residual
- Test significance of coefficients-T-statistic
- Normal Distribution

<u>GLMs</u>

- Link functions
- Use dummy coding for categorical variables
- Deviance
- Test significance of coefficients-T-statistic
- Exponential family of distributions

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
- Fox, J., *An R and S-PLUS Companion to Applied Regression*, Sage Publications