Introduction to Generalized Linear Models

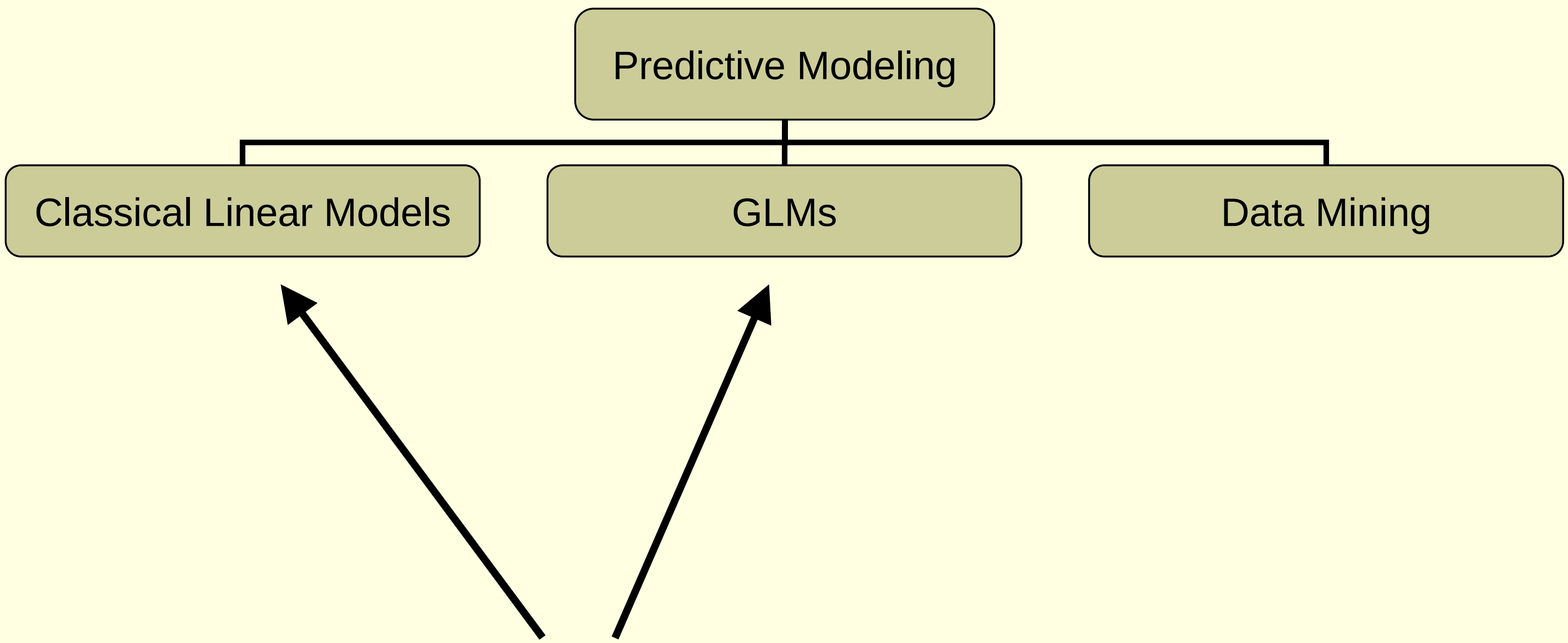
2007 CAS Predictive Modeling Seminar

Prepared by
Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
www.data-mines.com
Louise_francis@msn.com
October 11, 2007

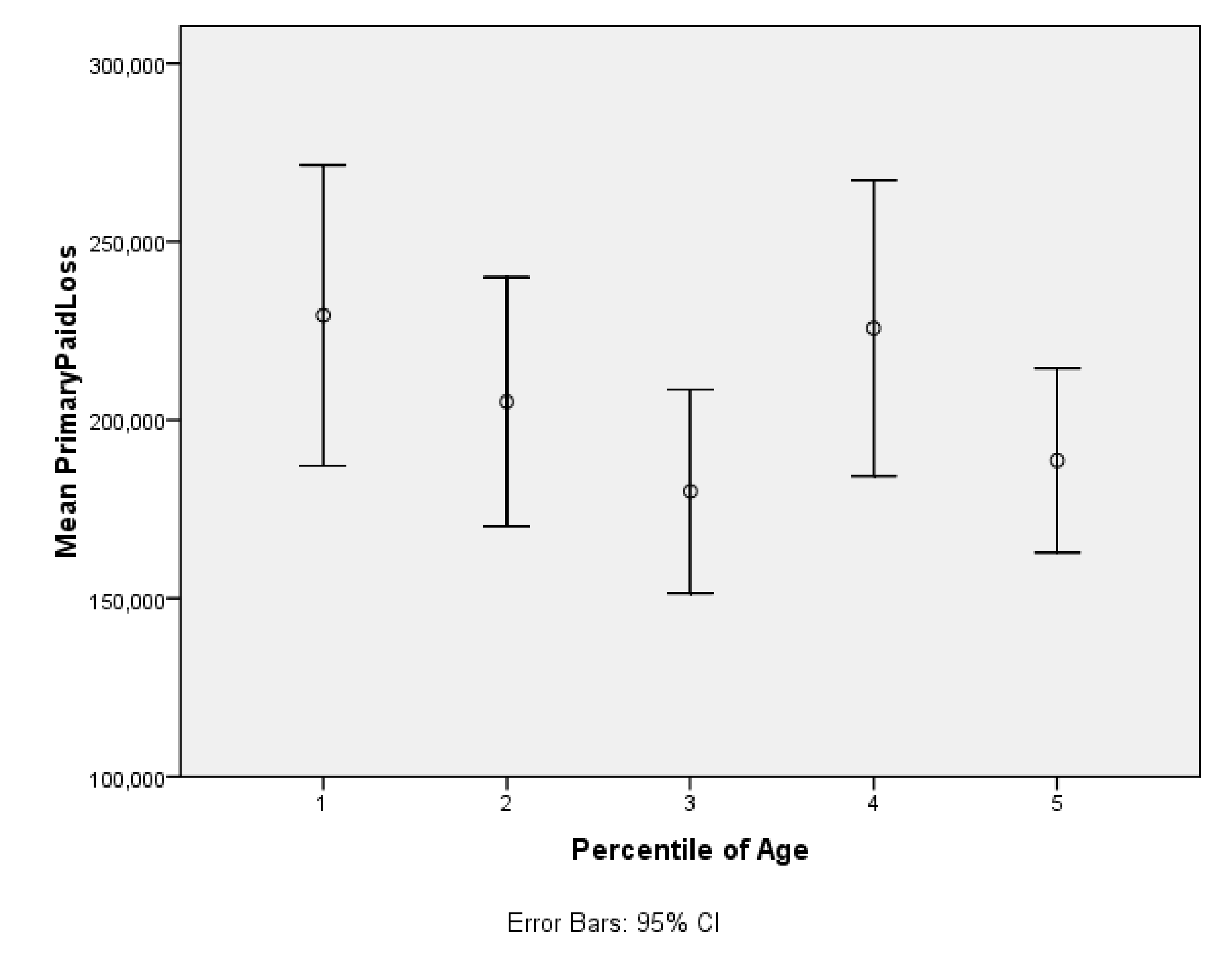
Objectives

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

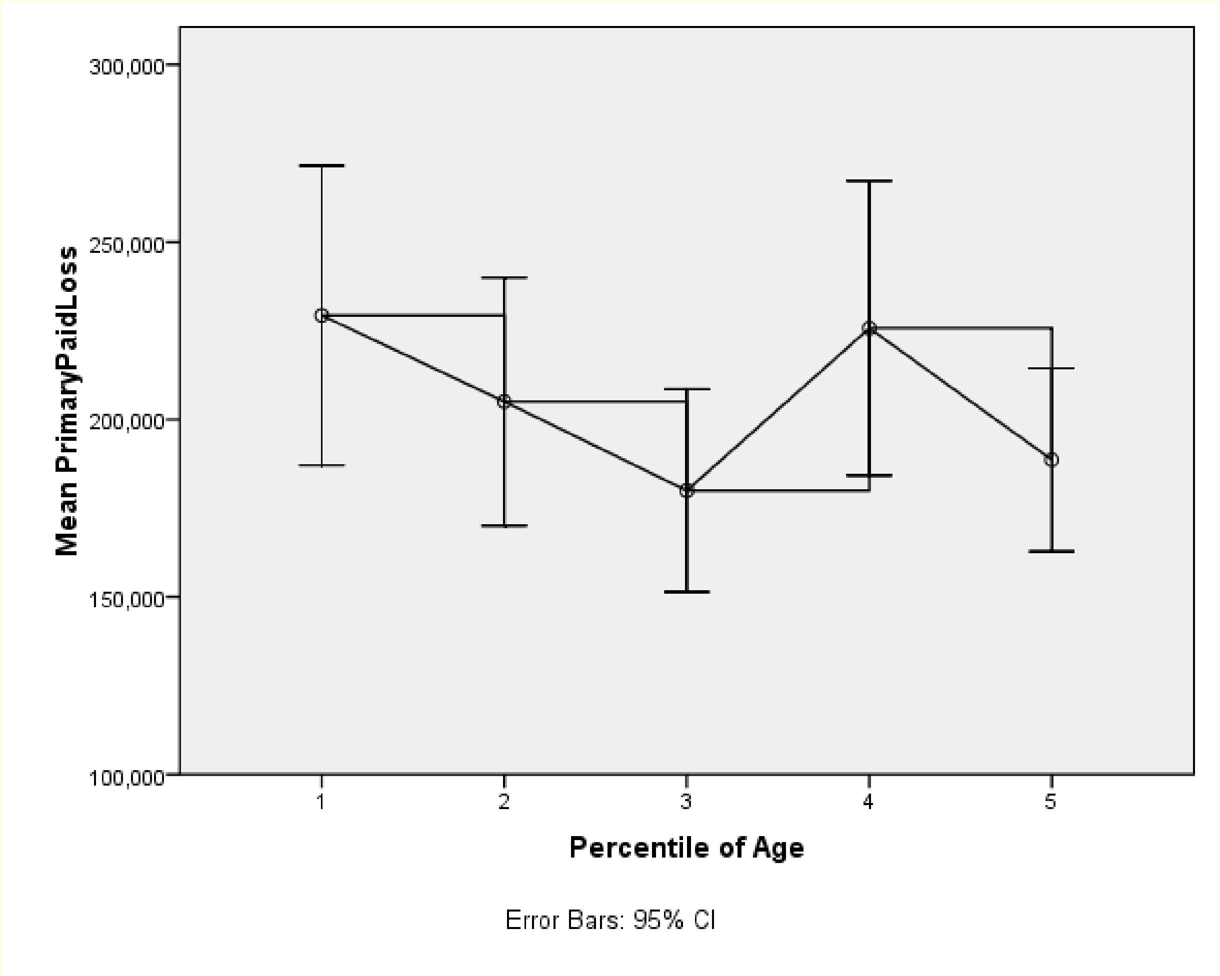
Predictive Modeling Family



Linear Models Are Basic Statistical Building Blocks: Ex:Mean Payment by Age Group

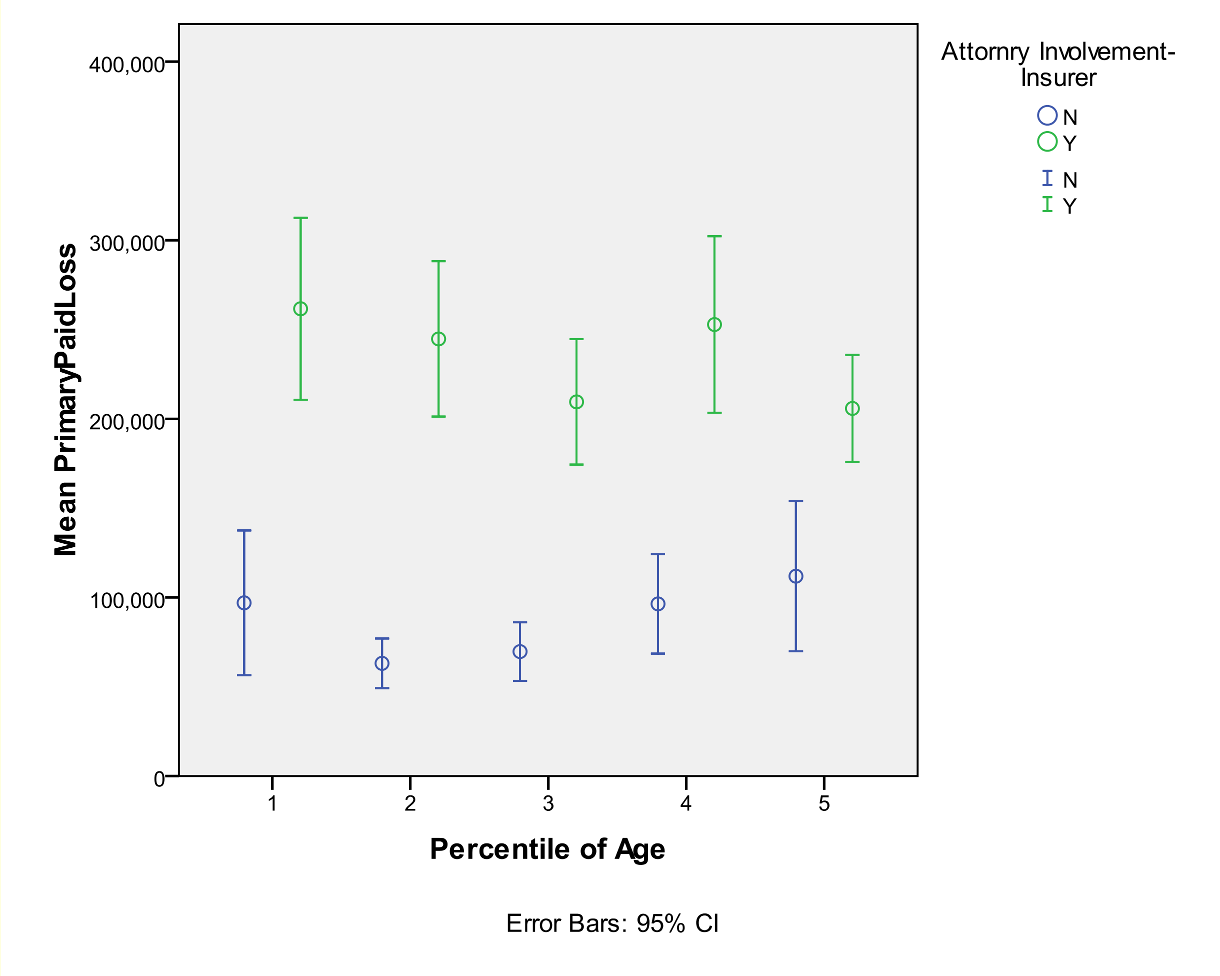


Linear Model for Means: A Step Function Ex:Mean Payment by Age Group

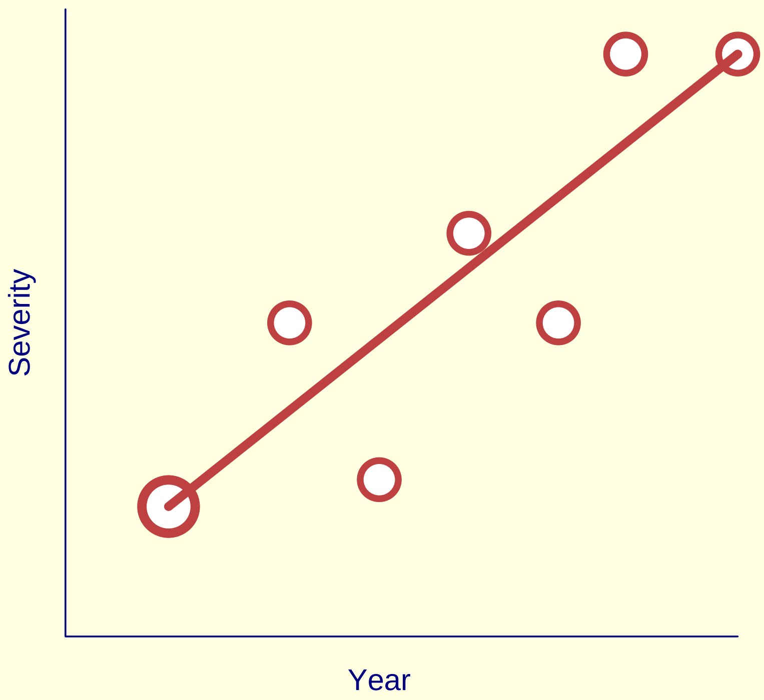


Linear Models Based on Means

Payment by Age Group and Attorney Involvement

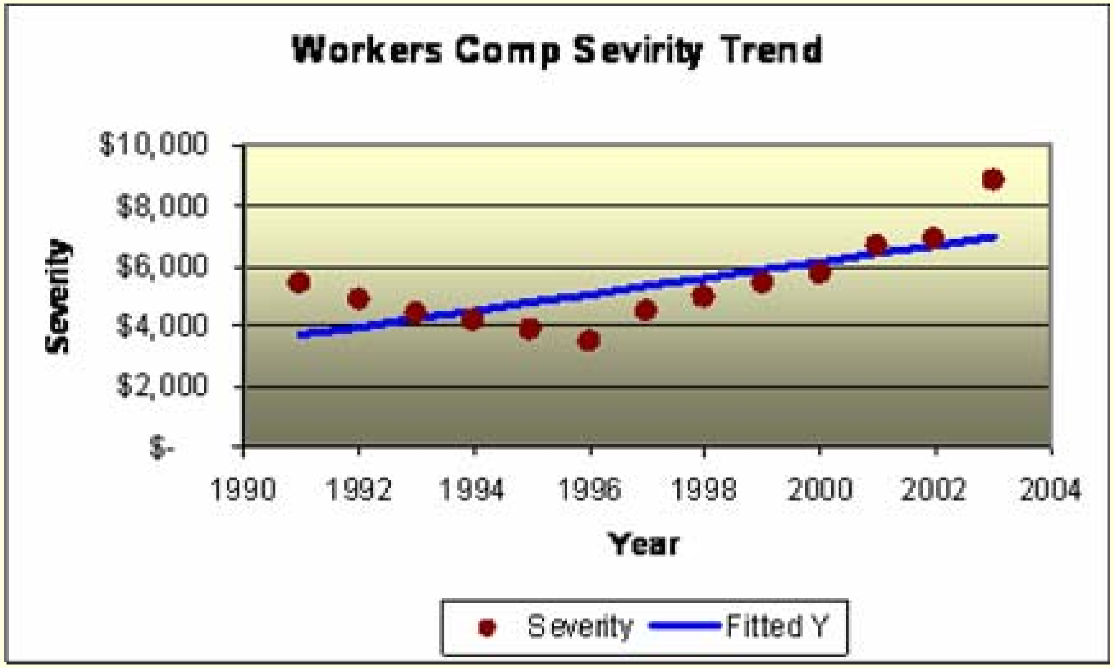


An Introduction to Linear Regression



Intro to Regression Cont.

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



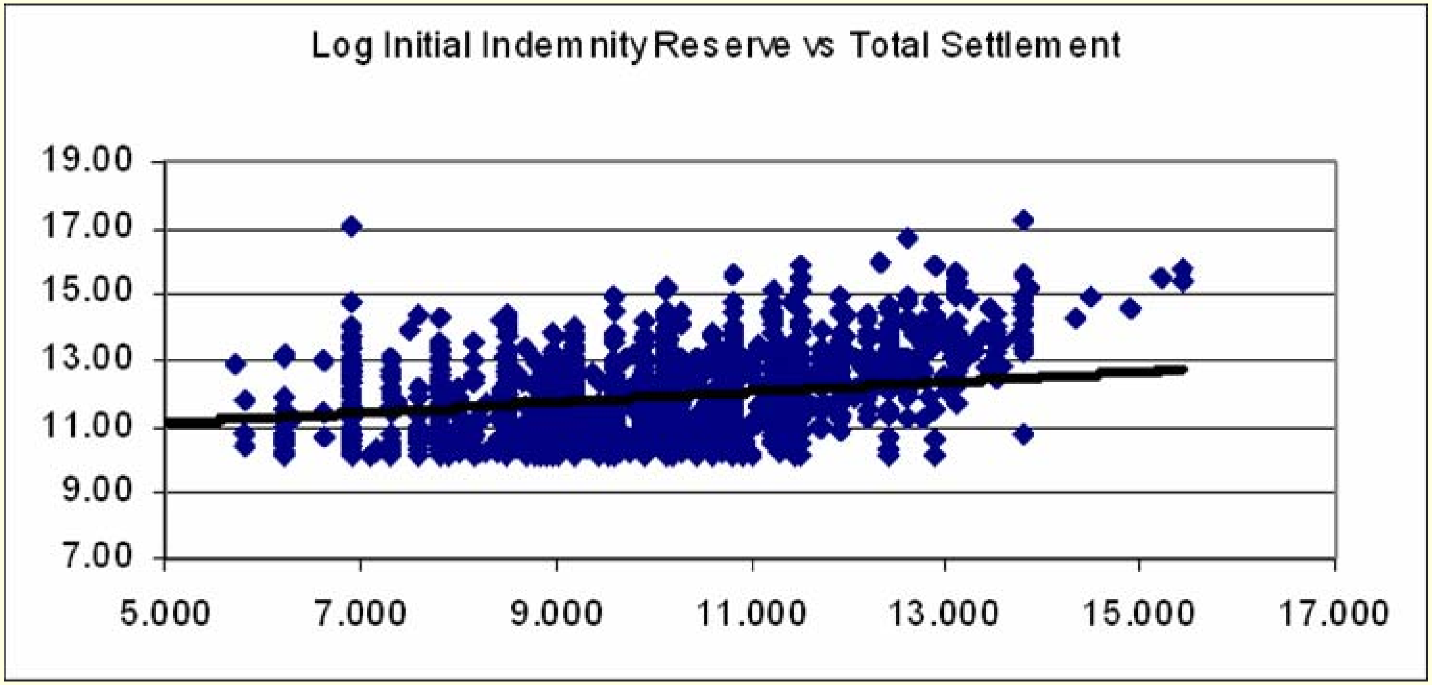
Some Work Related Liability Data

Closed Claims from Tx Dept of Insurance

- Total Award
- Initial Indemnity reserve
- Policy Limit
- Attorney Involvement
- Lags
  - Closing
  - Report
- Injury
  - Sprain, back injury, death, etc
- Data, along with some of analysis will be posted on internet

Simple Illustration

Total Settlement vs. Initial Indemnity Reserve

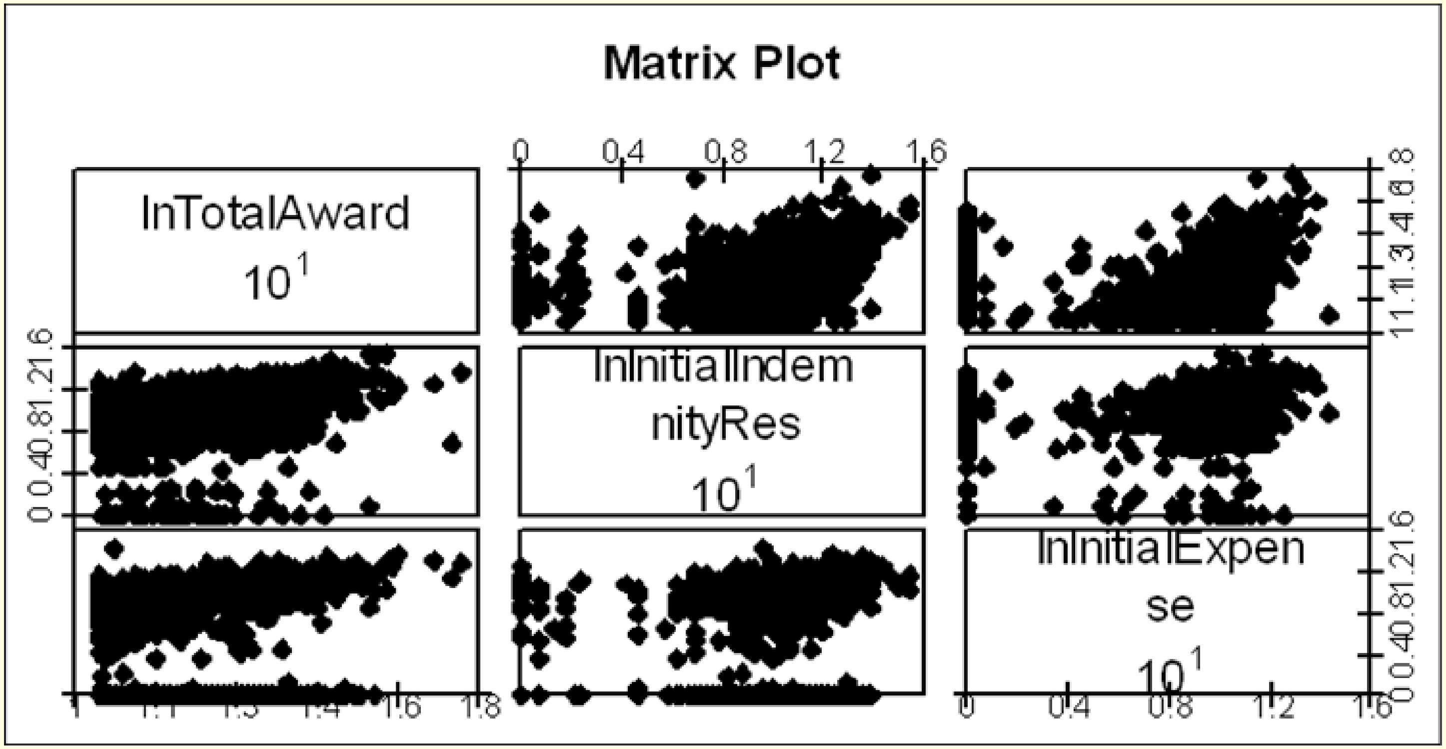


How Strong Is Linear Relationship?: Correlation Coefficient

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

	lnInitialIndemnityRes	*lnTotalAward*	*lnInitialExpense*	*lnReportlag*
lnInitialIndemnityRes	1.000			
lnTotalAward	0.303	1.000		
lnInitialExpense	0.118	0.227	1.000	
lnReportlag	-0.112	0.048	0.090	1.000

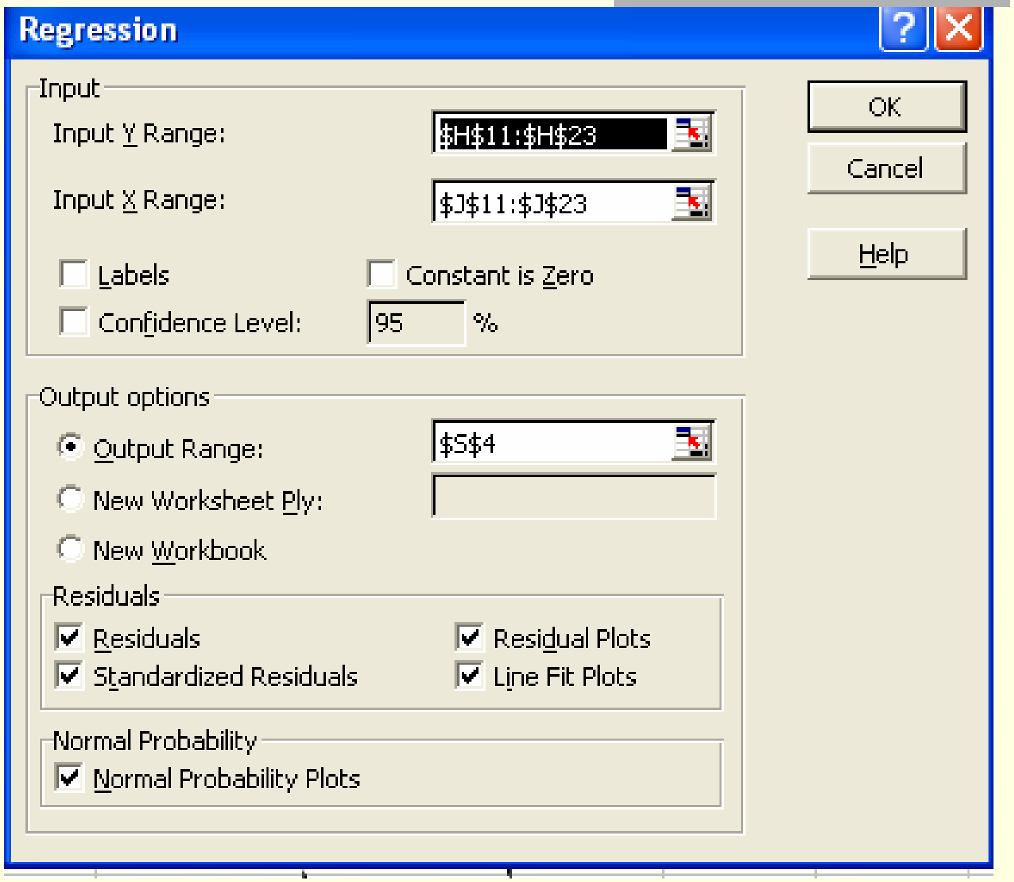
Scatterplot Matrix



Prepared with Excel add-in XLMiner

Linear Modeling Tools Widely Available: Excel Analysis Toolpak

- 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	0.303				
R Square	0.092				
Adjusted R Square	0.091				
Standard Error	1.206				
Observations	1818				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	266.29	266.29	183.07	0.00
Residual	1816	2641.50	1.45		
Total	1817	2907.79			

First Step: Compute residual

- Residual = actual – fitted

Y=lnTotal Award	Predicted	Residual
10.13	11.76	-1.63
14.08	12.47	1.61
10.31	11.65	-1.34

- 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 is Mean Square Error

R^2 Statistic

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.3757
R Square	0.1412
Adjusted R Square	0.1388
Standard Error	1.1740
Observations	1818

Significance of Regression

- F statistic:
  - (Mean square error of Regression/Mean Square Error of Residual)
  - Df of numerator = k = number of predictor vars
  - Df denominator = N - k

ANOVA (Analysis of Variance) Table

- Standard way to evaluate fit of model
- Breaks Sum Squared Error into model and residual components

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	410.5	82.1	59.6	0.00
Residual	1812	2497.3	1.4		
Total	1817	2907.8			

Goodness of Fit Statistics

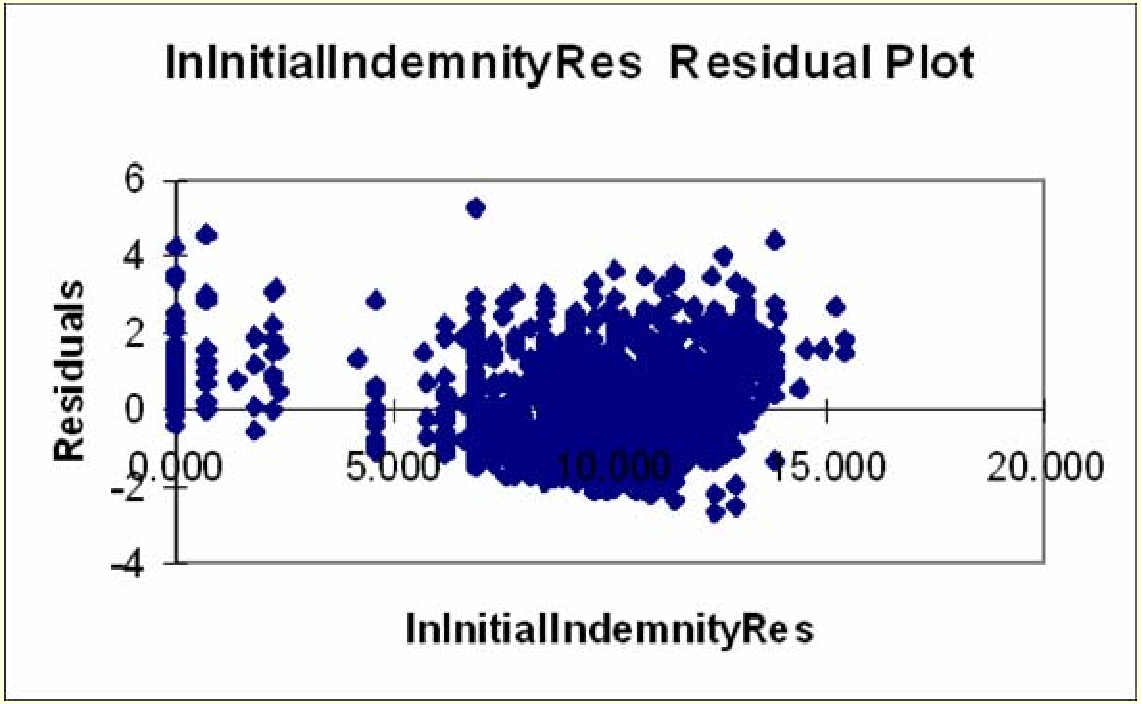
- T statistics: Uses SE of coefficient to determine if it is significant
  - SE of coefficient is a function of ***s*** (standard error of regression)
  - Uses T-distribution for test
  - It is customary to drop variable if coefficient not significant

T-Statistic: Are the Intercept and Coefficient Significant?

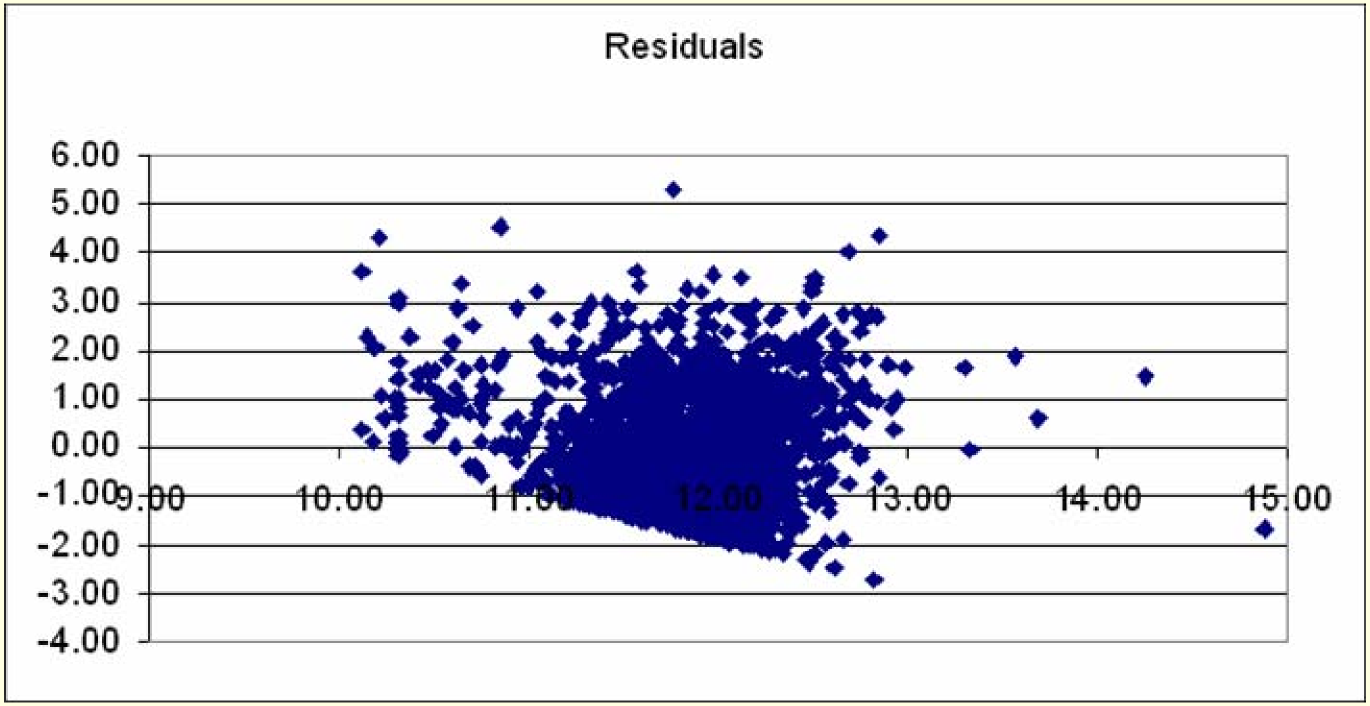
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	10.343	0.112	92.122	0
lnInitialIndemnity Res	0.154	0.011	13.530	8.21E-40

Other Diagnostics: Residual Plot Independent Variable vs. Residual

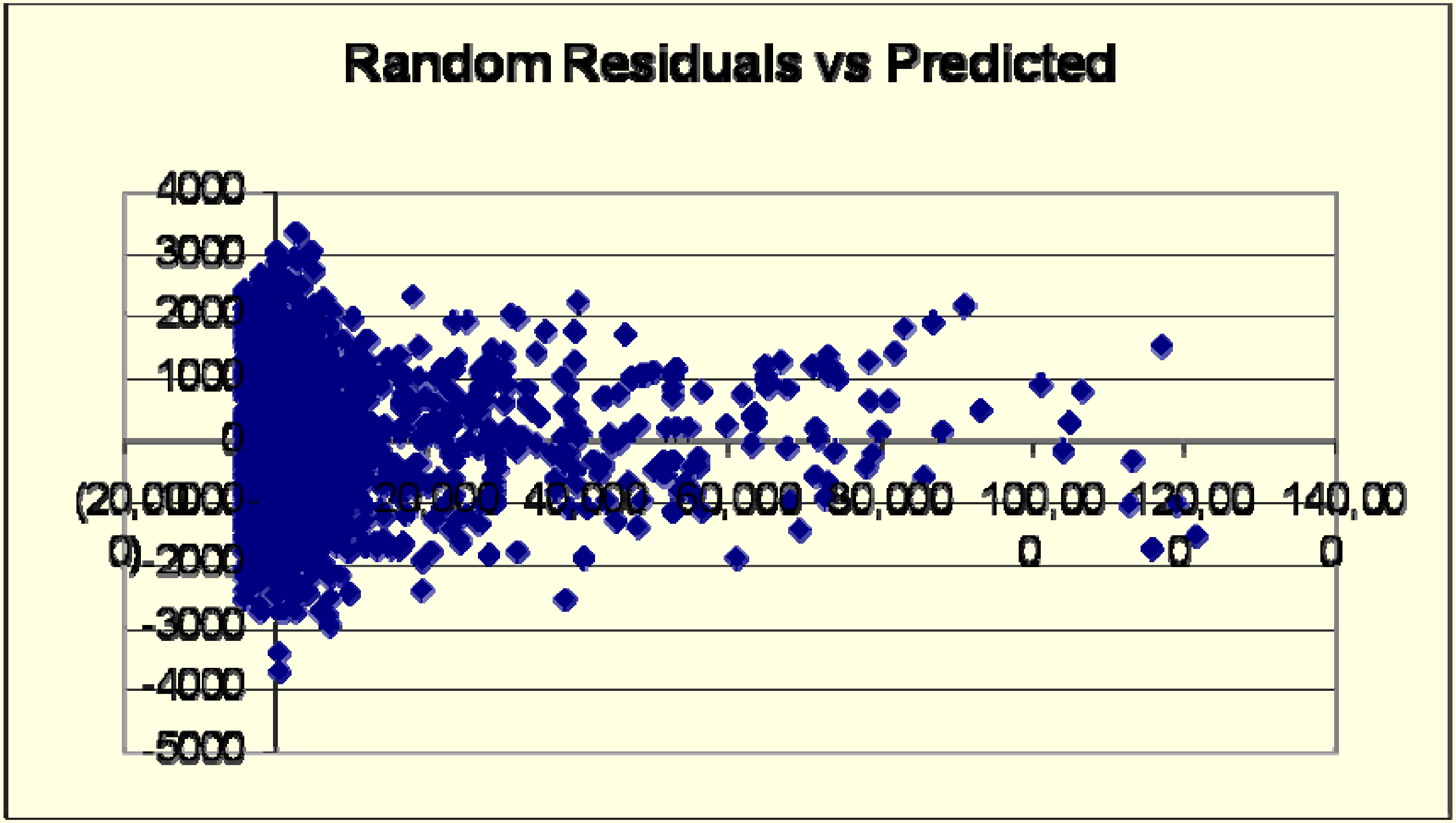
- Points should scatter randomly around zero
- If not, regression assumptions are violated



Predicted vs. Residual



Random Residual



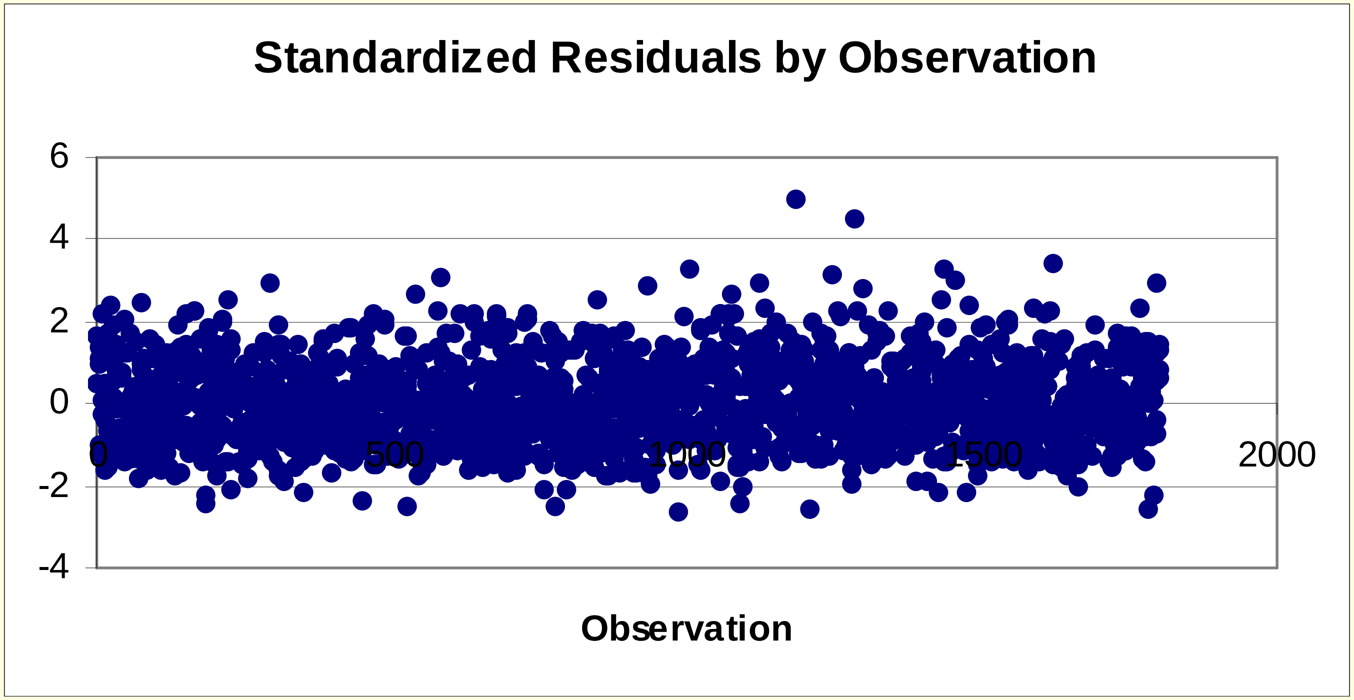
DATA WITH NORMALLY DISTRIBUTED ERRORS RANDOMLY GENERATED

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
  - May need to use weighted regression
- May indicate a nonlinear relationship

Standardized Residual: Find Outliers

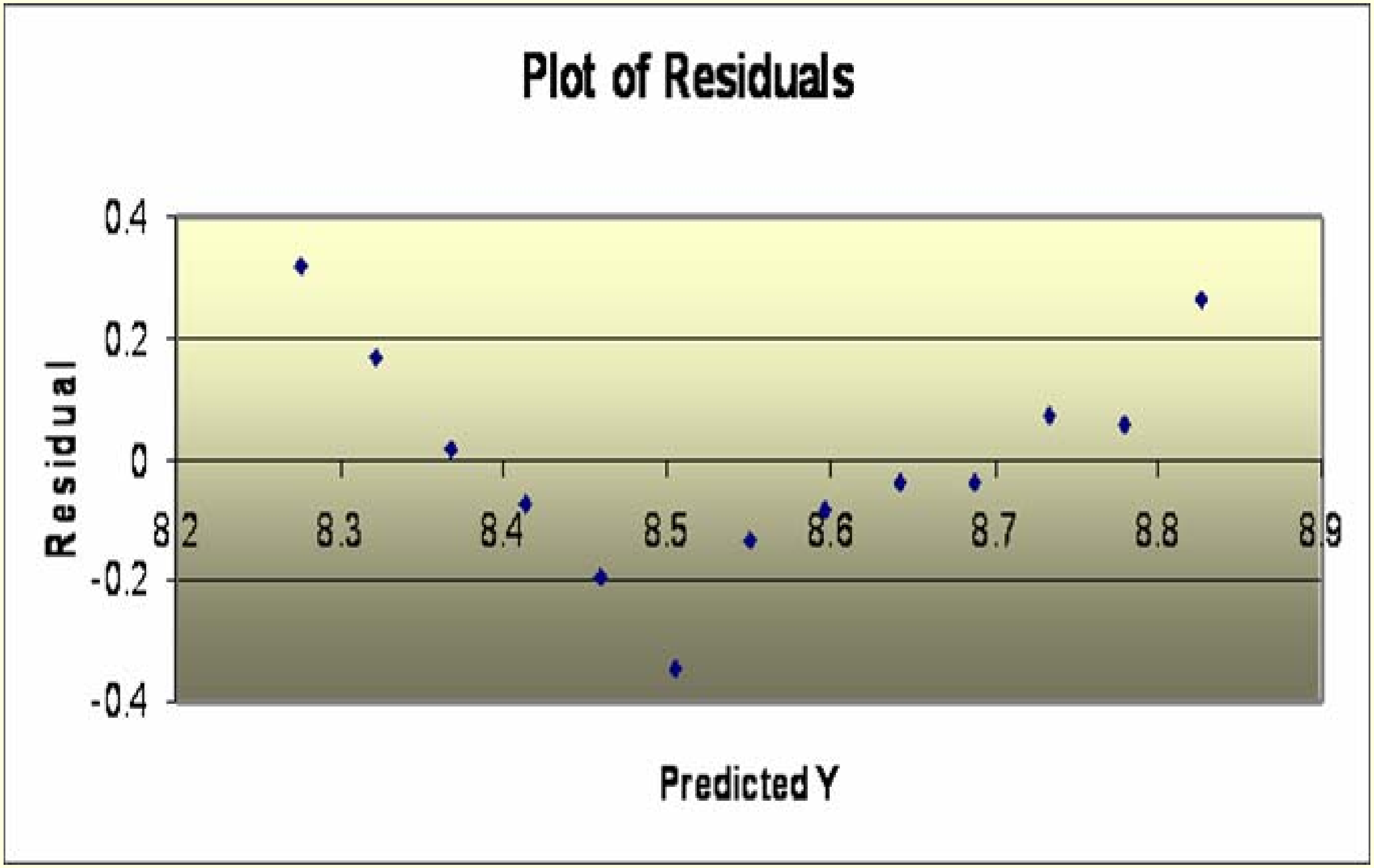
$$z_i = \frac{(y_i - \hat{y}_i)}{\sigma_{se}}, \quad \sigma_{se} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N - k - 1}}$$



Outliers

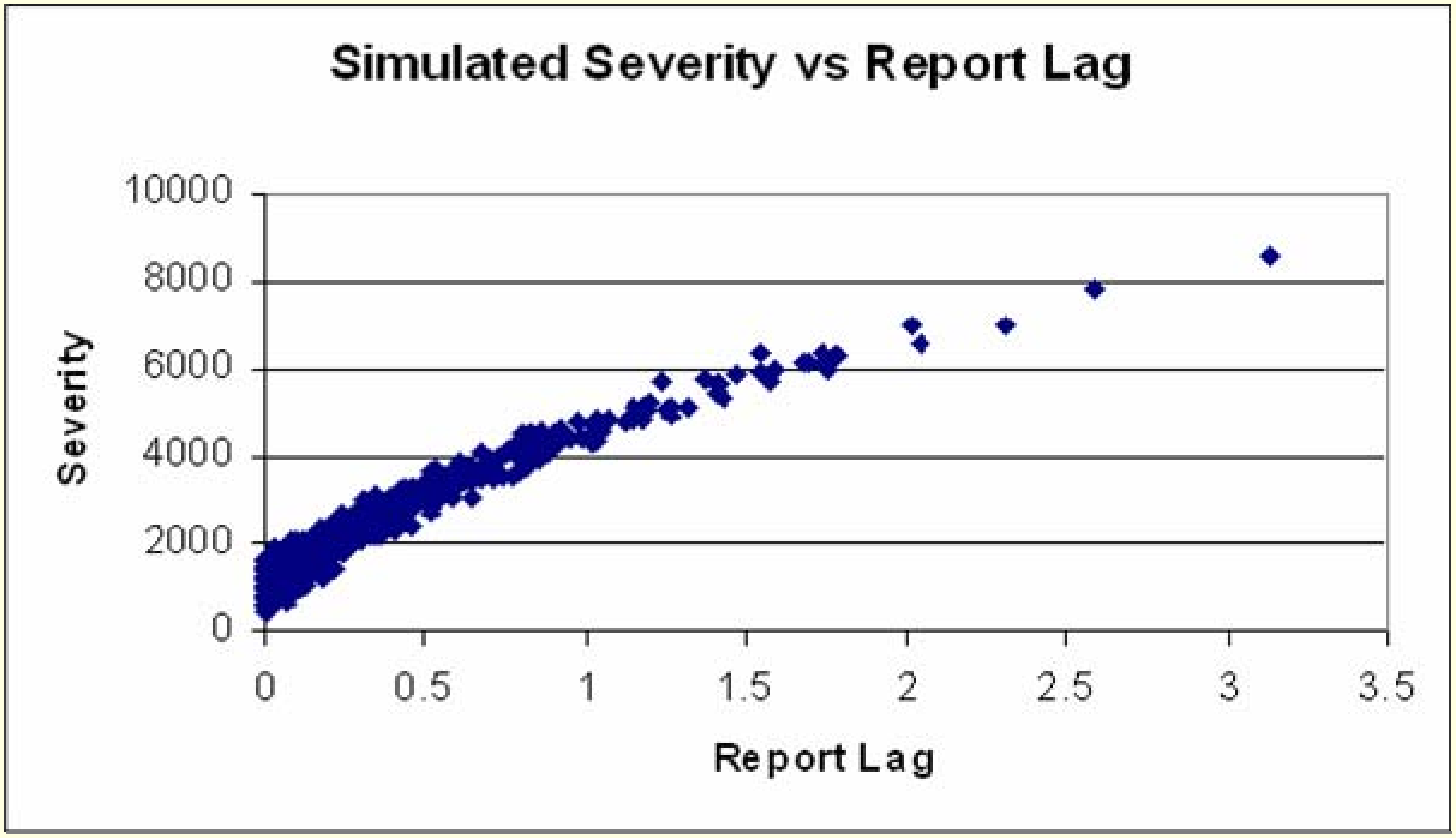
- May represent error
- May be legitimate but have undue influence on regression
- Can downweight oultliers
  - Weight inversely proportional to variance of observation
  - Robust Regression
    - Based on absolute deviations
    - Based on lower weights for more extreme observations

Non-Linear Relationship



Non-Linear Relationships

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

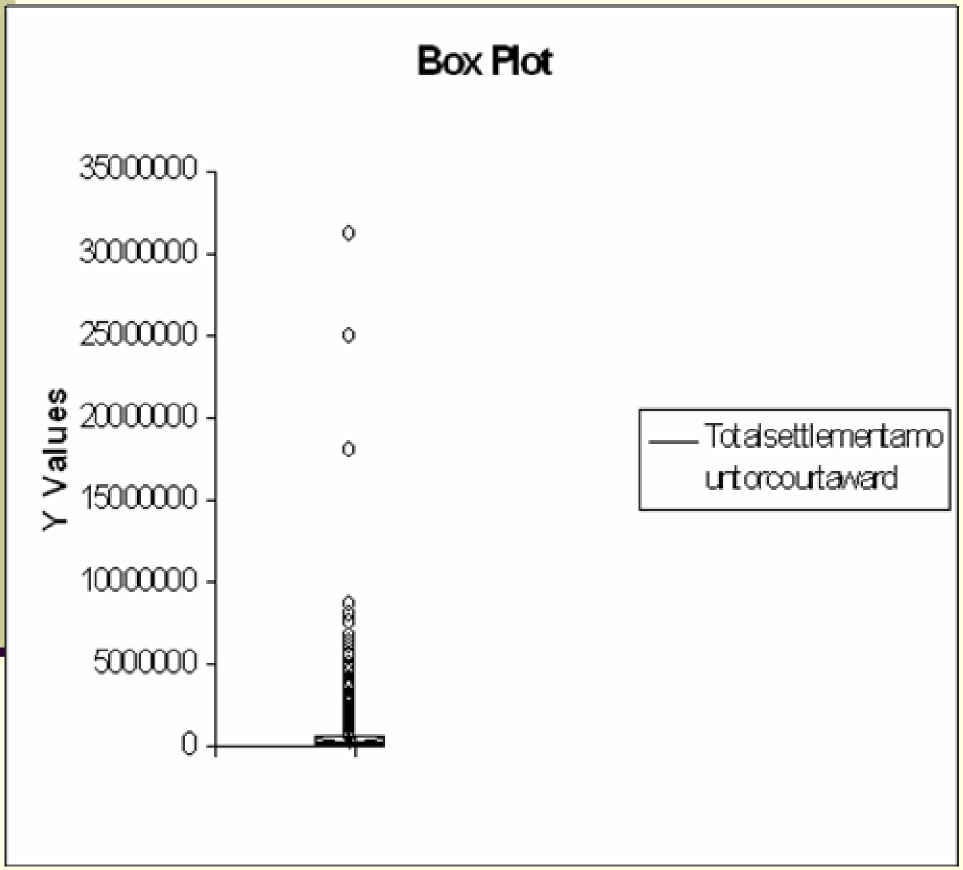


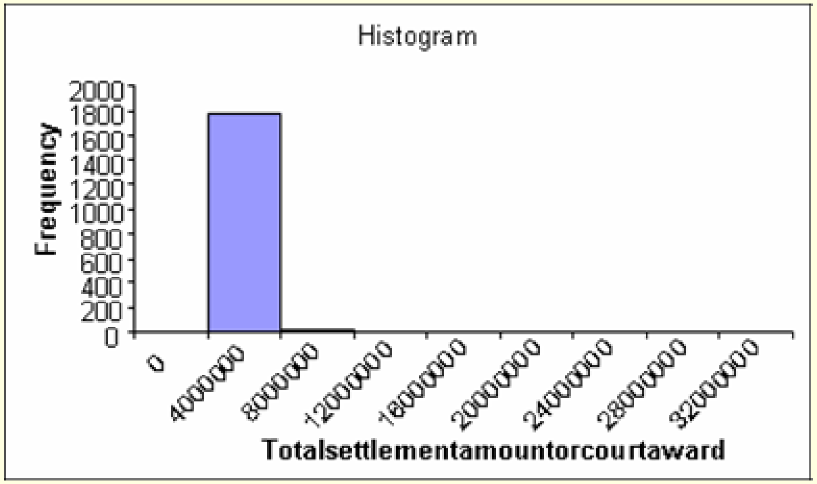
Transformation of Variables

- 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: Skewness of Distribution

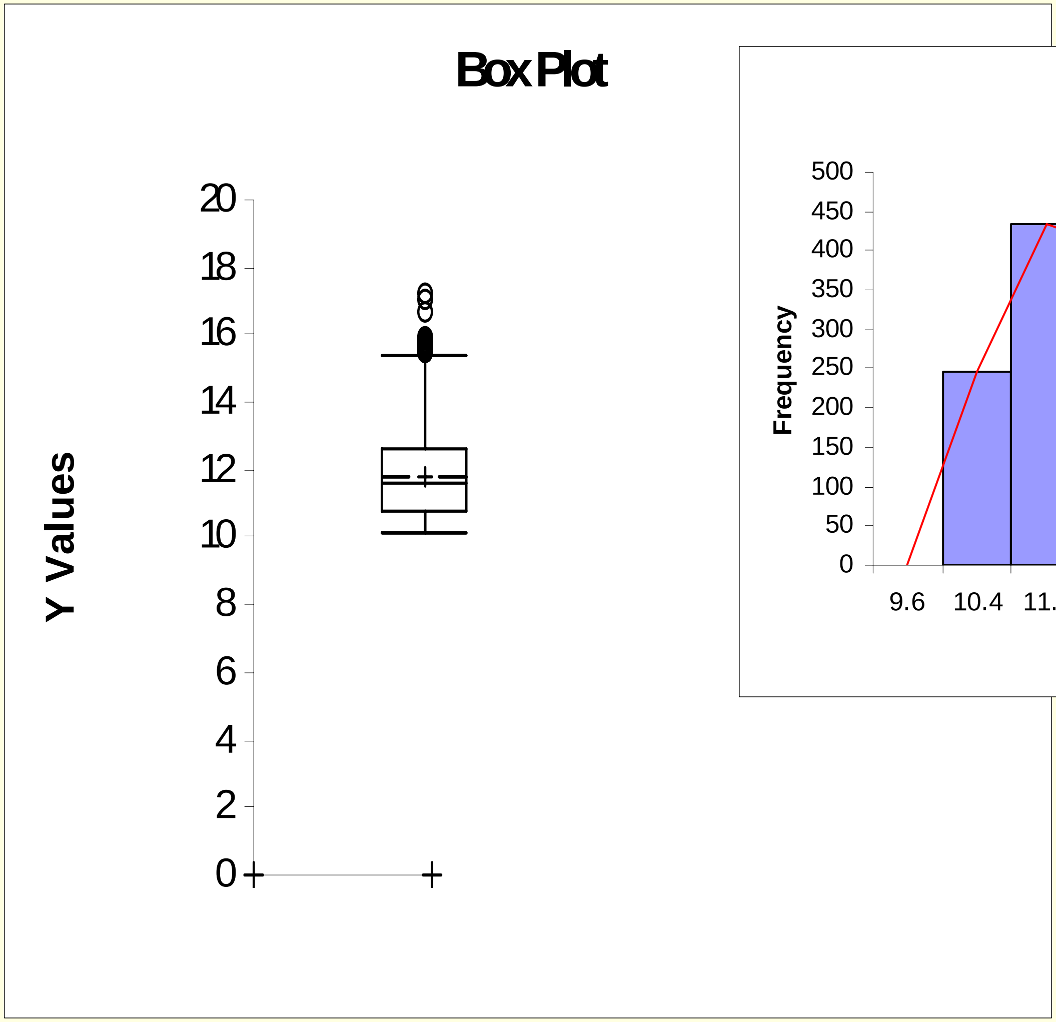
Use Exploratory Data Analysis to detect skewness, and heavy tails

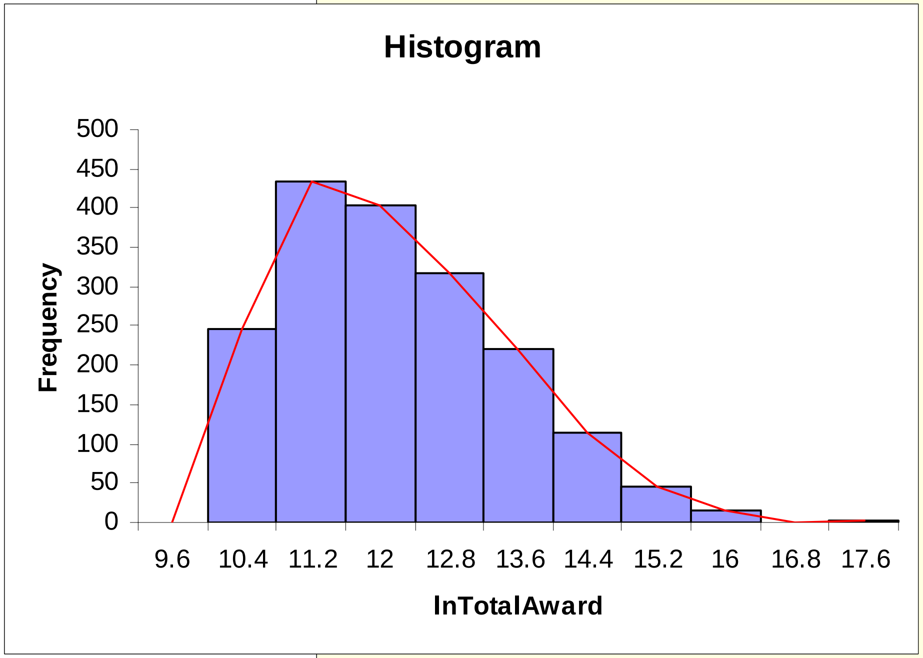




After Log Transformation

-Data much less skewed, more like Normal, though still skewed





Transformation of Variables

- Suppose the Claim 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

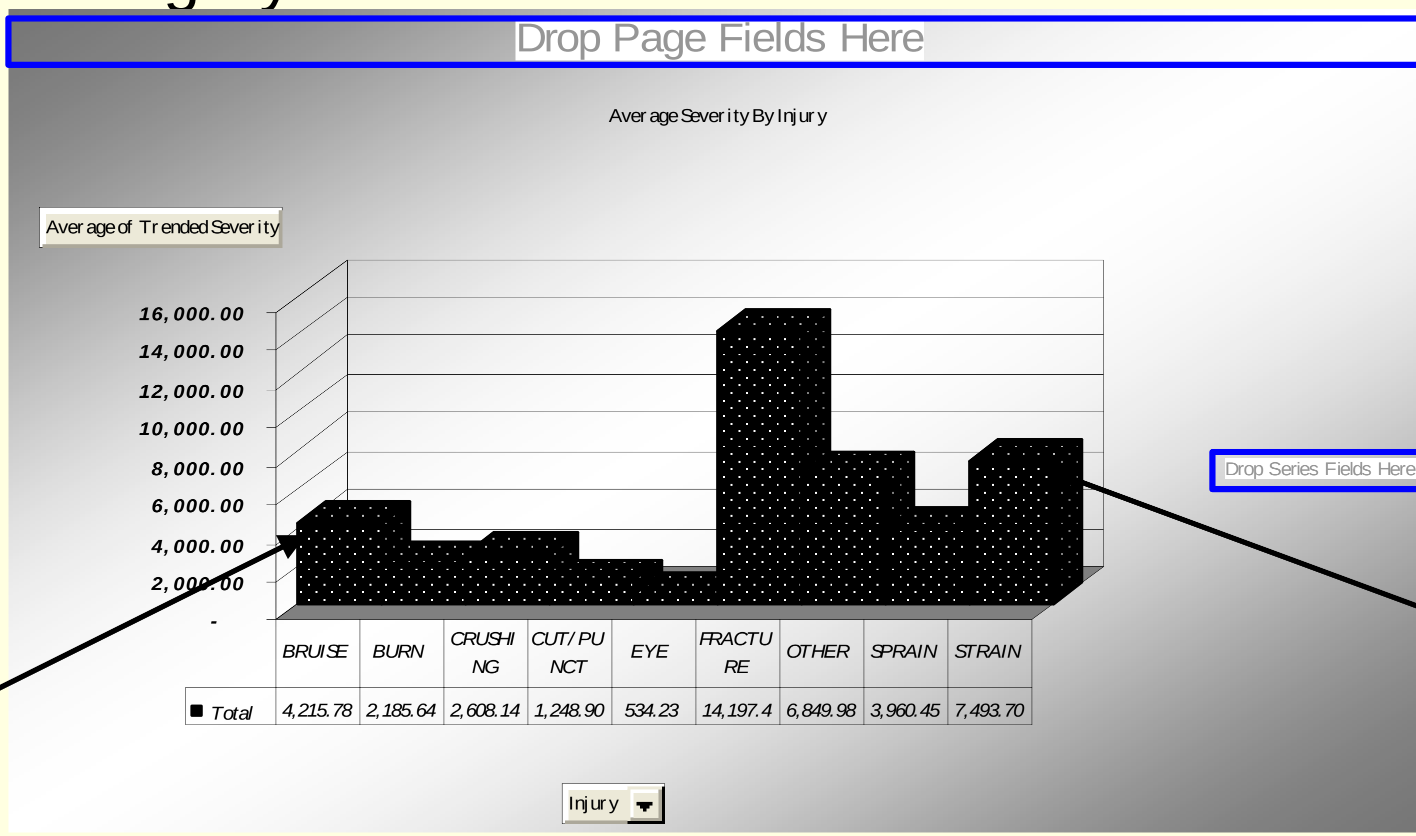
Categorical Independent Variables: The Other Linear Model: ANOVA

Average of Totalsettlementamountorcourtaward	
Injury	Total
Amputation	567,889
Backinjury	168,747
Braindamage	863,485
Burnschemical	1,097,402
Burnsheat	801,748
Circulatorycondition	302,500

Table above created with Excel Pivot Tables

Model

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



$Y=a_1$

$Y=a$

Two Categories

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

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

	Variable 1	*Variable 2*
Mean	124,002	440,758
Variance	2.35142E+11	1.86746E+12
Observations	354	1448
Hypothesized Mea	0	
df	1591	
t Stat	-7.17	
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	

Use T-Test from Excel Data Analysis Toolpak

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 Attorney Involvement
- Variable is 1 If Attorney Involved, and 0 Otherwise

Attorneyinvolvement-insurer	Attorney	TotalSettlement
Y	1	25000
Y	1	1300000
Y	1	30000
N	0	42500
Y	1	25000
N	0	30000
Y	1	36963
Y	1	145000
N	0	875000

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
782.08	FRACTURE	0	0	0	0	1	0	0	0
798.75	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

Injury Code	Injury_Backinjury	Injury_Multipleinjuries	Injury_Nervouscondition	Injury_Other
1	0	0	0	0
1	0	0	0	0
12	1	0	0	0
11	0	1	0	0
17	0	0	0	1

Top table Dummy variables were hand coded, Bottom table dummy variables created by XLMiner.

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

Multiple Regression

$Y = a + b1 * \text{Initial Reserve} + b2 * \text{Report Lag} + b3 * \text{PolLimit} + b4 * \text{age} + c_i \text{Attorney}_i + d_k \text{Injury}_k + e$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.49844
R Square	0.24844
Adjusted R Square	0.24213
Standard Error	1.10306
Observations	1802

ANOVA

	df	*SS*	*MS*	*F*
Regression	15	718.36	47.89	39.360
Residual	1786	2173.09	1.22	
Total	1801	2891.45		

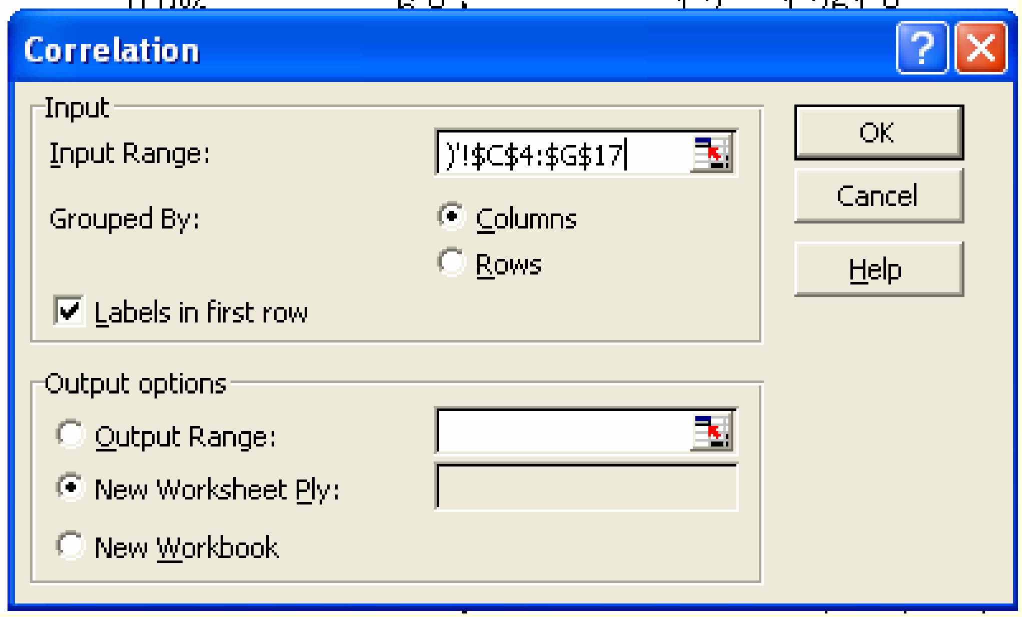
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	10.052	0.156	64.374	0.000
lnInitialIndemnityRes	0.105	0.011	9.588	0.000
lnReportlag	0.020	0.011	1.887	0.059
Policy Limit	0.000	0.000	4.405	0.000
Clmt Age	-0.002	0.002	-1.037	0.300
Attorney	0.718	0.068	10.599	0.000
Injury_Backinjury	-0.150	0.075	-1.995	0.046
Injury_Braindamage	0.834	0.224	3.719	0.000
Injury_Burnschemical	0.587	0.247	2.375	0.018
Injury_Burnsheat	0.637	0.175	3.645	0.000
Injury_Circulatorycondition	0.935	0.782	1.196	0.232

More Than One Categorical Variable

- For each categorical variable
  - Create k-1 Dummy variables
  - K is the total number of variables
  - 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

Ins Index	CPI	Employment	PchangeEmp	UEP Rate	Cng UEP	Residual	Resi
11.7	136.2	117,718					
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

Similarities with GLMs

Linear Models

- Transformation of Variables
- Use dummy coding for categorical variables
- Residual
- Test significance of coefficients

GLMs

- Link functions
- Use dummy coding for categorical variables
- Deviance
- Test significance of coefficients

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
- *Data Mining for Business Intelligence, Concepts, Applications and Techniques in Microsoft Office Excel with XLMiner,* Shmueli, Patel and Bruce, Wiley 2007
- De Jong and Heller, Generalized Linear Models for Insurance Data, Cambridge, 2008