

## Models for Data Analysis and Forecasting

Growing availability of on-line data

x **Powerful software tools**

= Expanding opportunities to extract  
valuable information from data mines

- **Regression Analysis**
  - ▶ the work-horse of data analysis and forecasting
  - ▶ software tools widely available

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Regression Analysis

- **Basic idea**
  - ▶ Model a so-called "dependent" variable,  $Y$ , as a function of some "independent" variable(s),  $X_1, X_2, \dots, X_k$ :
- $Y = f(X_1, X_2, \dots, X_k) + u$
- **Uses of the model:**
  - ▶ explain changes in  $Y$  from changes in  $X_1, X_2, \dots, X_k$
  - ▶ predict the value of  $Y$ , given values of  $X_1, X_2, \dots, X_k$
  - ▶ control the value of  $Y$  by setting values of  $X_1, X_2, \dots, X_k$

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Regression Analysis - the simple linear model

- Important particular case
  - ▶ simple linear regression
    - $Y = a + bX + u$
  - ▶ multiple linear regression
    - $Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + u$
- The approach
  - ▶ assume the relationship exists
  - ▶ identify the 'best' estimates of  $a, b_1, b_2, \dots, b_k$  based on the available data sample
- Larger sample data  $\Rightarrow$  Better estimates

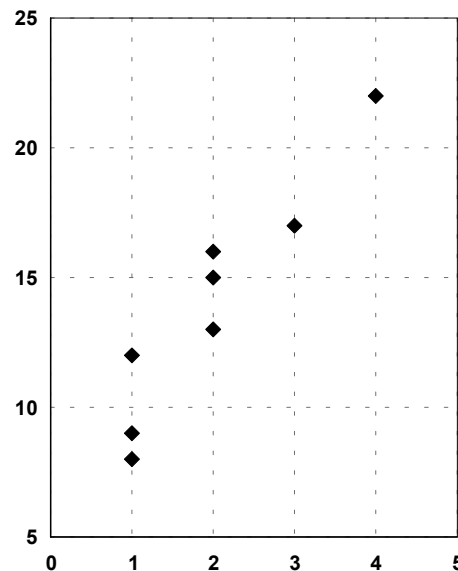
Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Basis: Minimize Sum of Squared Errors

- Question: How to construct a line such that the "error" of the line's fit to the data is minimized?

$x$	$y$
2	16
1	9
3	17
1	12
4	22
2	13
1	8
2	15



Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Classical Optimization Problem, With A Small Twist

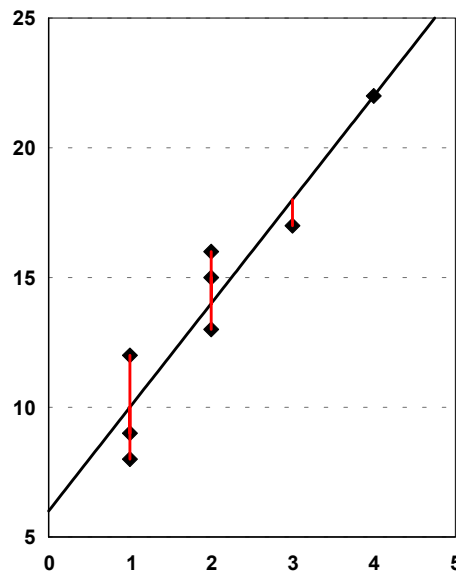
- Assume:  $y_{\text{est}} = \alpha + \beta x$
- Then, the "error" of making this estimate is
  - ▶ error =  $y - y_{\text{est}}$
  - ▶ error =  $y - \alpha - \beta x$
- A Typical Notion of Error Minimizing is to Minimize the Sum of the Errors Squared, e.g.;
  - ▶  $\min \sum (y - \alpha - \beta x)^2$
- Taking the First Derivative with respect to the coefficients  $\alpha$  and  $\beta$  gives equations allowing one to solve them

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## First Order Conditions

- $2\sum (y - \alpha - \beta x) = 0$   
(derivative w.r.t  $\alpha$ )
  - ▶ (remembering that  $\sum y = n y_{\text{avg}}$ )
  - ▶  $\alpha = y_{\text{avg}} - \beta x_{\text{avg}}$
- $2\sum x(y - \alpha - \beta x) = 0$ 
  - ▶  $b = \frac{(\sum xy - n x_{\text{avg}} y_{\text{avg}})}{(\sum x^2 - n (x_{\text{avg}})^2)}$



Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

### Using the Data Set....

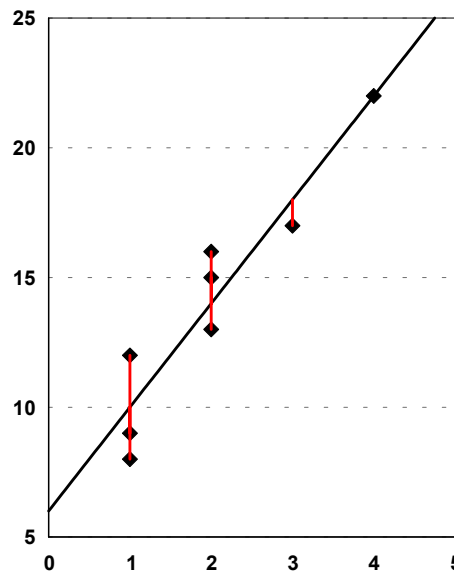
- Therefore
  - $\beta = 4$
  - $\alpha = 6$

- Minimizes the error you get by assuming that a line fits the data

	x	y	xy	x <sup>2</sup>
	2	16	32	4
	1	9	9	1
	3	17	51	9
	1	12	12	1
	4	22	88	16
	2	13	26	4
	1	8	8	1
	2	15	30	4
<b>Sum</b>	<b>16</b>	<b>112</b>	<b>256</b>	<b>40</b>
<b>averages</b>	<b>2</b>	<b>14</b>		
<b>n =</b>	<b>8</b>			

### So, How Well Did We Do?

- Looks Good
- But, How Good?
- We Need To Generate Some Statistics To Give Us An Indication of the Quality of the Estimated Line
- Fall Back Upon What We Know



## We Minimized The Sum of Squared Errors

- But We Didn't Make That Sum = 0, as the Chart Shows.
- So, What Is The Sum of Squared Errors (SSE)?
  - ▶ What would the SSE be if we just assumed  $y = y_{avg}$ ?

	x	y	est y	(y - y est)^2	(y - y avg)^2
	2	16	14	4	4
	1	9	10	1	25
	3	17	18	1	9
	1	12	10	4	4
	4	22	22	0	64
	2	13	14	1	1
	1	8	10	4	36
	2	15	14	1	1
Avg		14			
Sum				16	144

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Measure Of Goodness of Fit

- We Define  $(y - y_{avg})_2$  to be the Total Sum of Squares (TSS)
- And the Sum of Squared Errors (SSE) Is the portion of the variation in y that our hypothesized line does NOT explain
- So, the percent of the TSS that our estimated line "explains" is just:
 
$$(TSS - SSE) \div TSS = (144 - 16) \div 144 = 89\%$$
- This Statistic Is Also Known as the  $R^2$  of the Regression
- We're Done With Derivations, But Note:
  - ▶ Multiple Linear Regression Works The Same Way, Just With Matrices
  - ▶ New Set of Issues Beyond  $R^2$  to consider

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Other Important Measures (But No More Derivations!)

- Standard Error of the Regression
  - ▶  $s^2 = (n-2)^{-1} ( \sum(y-y_{avg})^2 - \beta \sum(x-x_{avg})(y-y_{avg}) )$
- Standard Error of the Coefficients
  - ▶  $s_{\alpha}^2 = s^2 ( n^{-1} + x_{avg}^2 \div \sum(x - x_{avg})^2 )$
  - ▶  $s_{\beta}^2 = s^2 \div \sum(x - x_{avg})^2$
- Standard Error of the Estimate  $Y_i = \alpha + \beta x_i$ 
  - ▶  $s_{Y_i}^2 = s^2 ( n^{-1} + (x_i - x_{avg})^2 \div \sum(x - x_{avg})^2 )$

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Why These Standard Errors?

- Tests of Confidence
  - ▶ It turns out that the distribution for the estimated parameters and the estimate itself are distributed according to something known as the student t-statistic
  - ▶ Thus, a statistical measure of the importance of a variable in the regression can be constructed
- Consider: if an independent variable (x) is NOT relevant to the dependent variable (y), then it's "true" coefficient  $\beta = 0$ 
  - *This is a null hypothesis -  $H_0$*
  - ▶ Yet, we get an estimated value of  $\beta \neq 0$  from our data
  - ▶ So, what's right? The null hypothesis or the non-zero  $\beta$ ?

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Can State The Answer According To The Degree of Confidence We Wish To Assert

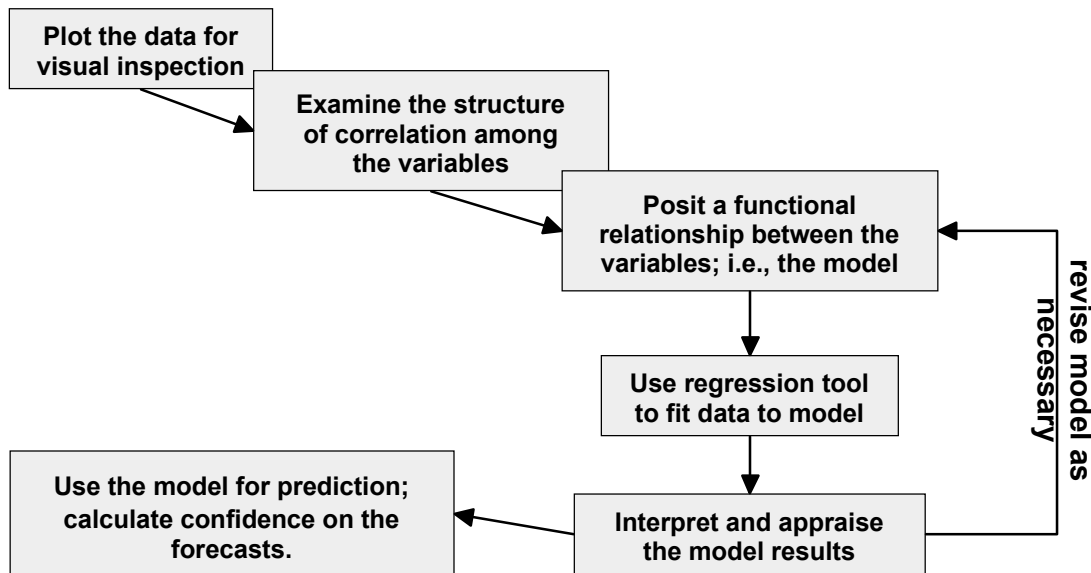
- If distributed according to  $t$ , then we mean that:
  - ▶  $(\beta_{est} - \beta) \div s_{\beta} \sim t$
- If  $H_0$  is False, then
  - ▶  $|\beta_{est} \div s_{\beta}| > t_{n-2}$  at the % desired
- This calculated term is what most regression packages return as the “t-statistic at NN% confidence”

n	Student's t-statistic			
	% of area under t-curve remaining			
	0.100	0.050	0.025	0.010
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681
13	1.350	1.771	2.160	2.650
14	1.345	1.761	2.145	2.624
15	1.341	1.753	2.131	2.602

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Regression Analysis - General Procedure



Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Regression Analysis With Excel

- **Scatter Plots:**
  - ▶ Go to 'Insert' menu
  - ▶ Chart...
  - ▶ XY (Scatter)
- **Correlation table: Go to 'Tools' menu**
  - ▶ Data Analysis...
  - ▶ Correlation
- **Regression:**
  - ▶ Go to 'Tools' menu
  - ▶ Data Analysis...
  - ▶ Regression

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Assessing the Regression Output

- **For the whole model**
  - ▶ **R-square** : measures the goodness of fit, the percentage of variation in Y that can be explained by the variation in the Xi's. The closer to 1, the better.
  - ▶ **Significance F** : a probability value that reflects the overall significance of the regression equation. A low value indicates high significance. To be confident enough that the relationship found between Y and the Xi's truly exists, i.e. is not due to chance, Significance F should be less than 0.05.
- **For each coefficient estimated**
  - ▶ **P-value** : a probability that reflects the significance of the coefficient. A low value indicates high significance. To be confident enough that this variable should be included in the model, P-value should be less than 0.05.

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Building Regression Models

- The issue:
  - ▶ which variables should be included in the multiple regression model?
  - ▶ in what form should they be included in the model?
- Model building = Art + Science
- No unique "best set" of explanatory variables
  - ▶ depends on purpose for which model is built:
    - *purely descriptive*
    - *prediction*
    - *control*
  - ▶ no fool-proof procedure to get the "best" model

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Building Regression Models

- General approaches
  - ▶ Fitting models to all possible subsets of variables not practical if many variables
- Inclusion or deletion of variables, one at a time
  - ▶ Forward Selection
  - ▶ Backward Elimination

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Backward Elimination

1. Fit model with all variables
2. Deal with multicollinearity if needed
3. Eliminate variable with lowest non-significant t-value
4. Fit model with remaining variables
5. Stop when all t-values are significant

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Multicollinearity

- a problem that may occur when some explanatory variables are strongly correlated
- the problem: confusing regression results, estimates of coefficients may be inaccurate

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**MSL**  
Materials Systems Laboratory

## Dealing With Multicollinearity

- Inspect correlation matrix for pairs of variables with corr coeff  $> .7$  or  $< -.7$
- Run regression model including both of the correlated variables
- Three cases may arise:
  - ▶ One of the two vars is not significant: then, drop it. Check to make sure the coeff of the one that stays makes sense
  - ▶ Both are not significant: drop one; check if the other becomes significant, sign of coeff?
    - *If no, drop both*
    - *Also, try swapping the two*
  - ▶ Both significant: inspect signs of coefficients (including intercept): do they make sense? If yes, can keep both.