Least Squares Curve Fitting

1st degree polynomial (line)

$y = a_0 + a_1x + e$

$e = y - a_0 - a_1x$

$e^2 = (y - a_0 - a_1x)^2$ ← The squared error

Minimizing the squared error we can obtain the best parameters for a_0 and a_1 using:

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

and

$a_0 = \bar{y} - a_1 \bar{x}$ where $\bar{x}$ represents the average of x and $\bar{y}$ represents the average of y

2nd degree polynomial (parabola)

$y = a_0 + a_1x + a_2x^2 + e$

Minimizing the squared error we can obtain the best parameters for a_0, a_1, and a_2 using:

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

M x X = N (need to solve for X matrix as follows)

Use Microsoft Excel you can invert the 3 x 3 matrix and multiply both sides by the inverted matrix on the left as shown below:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

X = M^{-1} x N

This will give you the optimal parameters for a_0, a_1, and a_2.

Linear Interpolation

When you have two data points (x_0, $f(x_0)$) and (x_1, $f(x_1)$) you can linearly interpolate between then to obtain an approximation for a value *x* such that $x_0 < x < x_1$, that is, the value you are looking for is between x_0 and x_1 using:

$$f_{est}(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

Example Problem

X	10	15	20	25	30
Y	35	50	60	77	95

Using the **4 non-shaded values** above, find a_0 and a_1 for the least squares linear regression. Compute the overall squared-error. Write the completed polynomial.

n = 4
Σ x = 75
Σ y = 240
Σ x Σ y = 18000
$\Sigma x^2 = 1625$

$(\Sigma x)^2 = 5625$
mean of x = 18.75
mean of y = 60
Σ x y = 5150

$$a1 = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

$$a1 = \frac{4(5150) - (75)(240)}{4(1625) - 5625}$$

a1 = 2.9714

a0 = mean of y – (a1)(mean of x)
a0 = 60.0 – (2.9714)(18.75)
a0 = 4.2857

Overall Squared Error $= \Sigma e^2 = \Sigma(y - ao - a1x)^2$
$= \Sigma (y_{actual} - y_{model})^2$
$= (35 - 4.2857 - 2.9714\,(10))^2 + ... + (95 - 4.2857 - 2.9714(30))^2$

Overall Squared Error = 18.5714

Least squares linear regression: Y = 4.2857 + 2.9714X

Using the **4 non-shaded values** above, find a0, a1, and a2 for a parabolic least squares regression (polynomial of degree 2). Use MS Excel to solve for these coefficients. Compute the overall squared-error. Write the completed polynomial.

	X	Y	X ^ 2	X x Y	X ^ 3	X ^ 4	X^2 * Y	Sq Err (1st)	Sq Err (2nd)
	10	35	100	350	1000	10000	3500	1.0000	0.6694
	15	50	225	750	3375	50625	11250	1.3061	4.7603
	20	60	400	1200	8000	160000	24000	13.7959	2.6777
	30	95	900	2850	27000	810000	85500	2.4694	0.0744
Sum	75	240	1625	5150	39375	1030625	124250	18.5714	8.1818
Mean	18.75	60							
Sum of X * Sum of Y	18000								
(Sum of X) ^ 2	5625								
a0	4.2857								
a1	2.9714								

4	75	1625	
75	1625	39375	**M**
1625	39375	1030625	

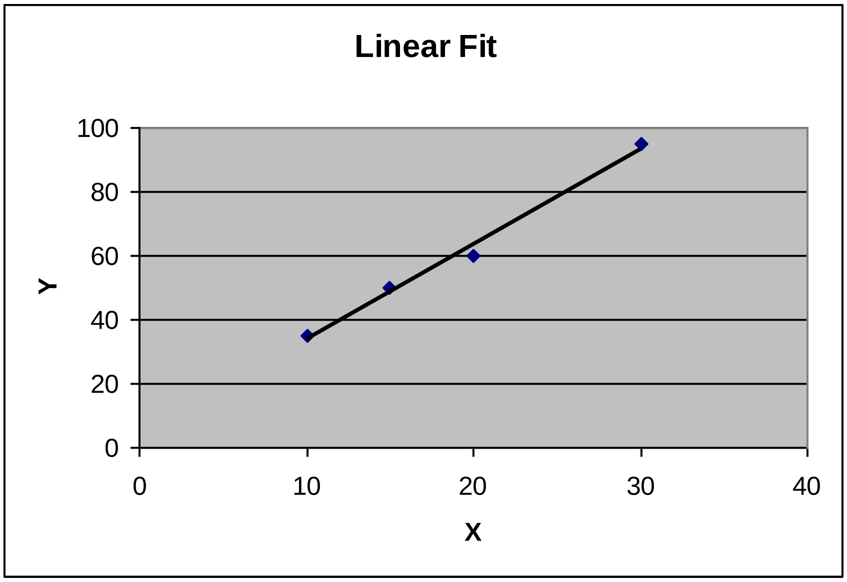
18.09091	-1.93636	0.045455	
-1.93636	0.215545	-0.00518	**M^{-1}**
0.045455	-0.00518	0.000127	

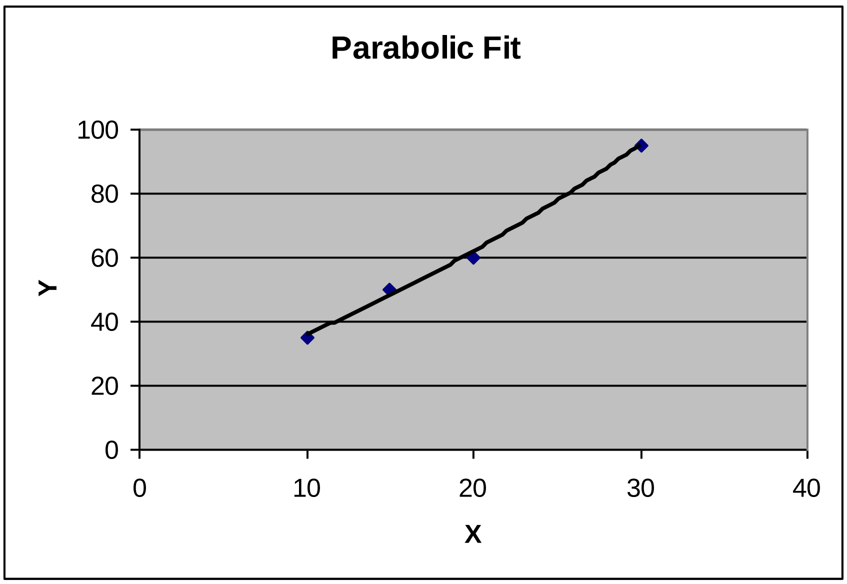
240	
5150	**N**
124250	

a0	17.27273	
a1	1.490909	**M^{-1} x N**
a2	0.036364	

Parabolic least squares regression: $Y = 17.27273 + 1.490909X + 0.036364X^2$

<u>NOTE:</u> IF YOU MISSED THE PROBLEM SOLVING INFORMATION ON COMPUTING MATRIX INVERSE (MINVERSE) AND MATRIX MULTIPLICATION (MMULT), SEE MICROSOFT EXCEL HELP FOR DETAILED INSTRUCTIONS AND EXAMPLES.





Fill in the following test table:

	Y values			Absolute Error: |Actual - Predicted|			Results		
X	linear interpolation	linear fit	parabolic fit	linear interpolation	linear fit	parabolic fit	actual	best value	best method
25	77.5	78.6	77.3	0.5	1.6	0.3	77	77.3	Parabolic Fit

Which method(s) performed the best? Would you have expected the outcomes? How do these perform for these data points vs. the linear and parabolic curve's squared errors? Discuss your answer.