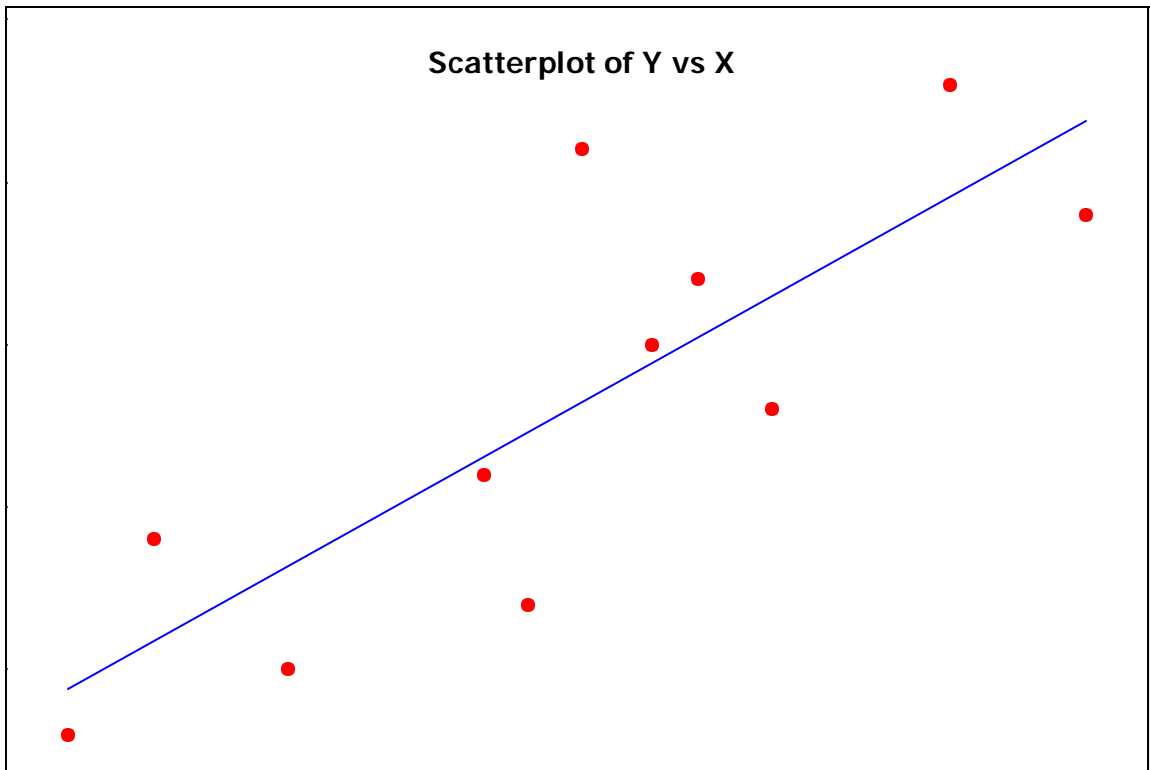


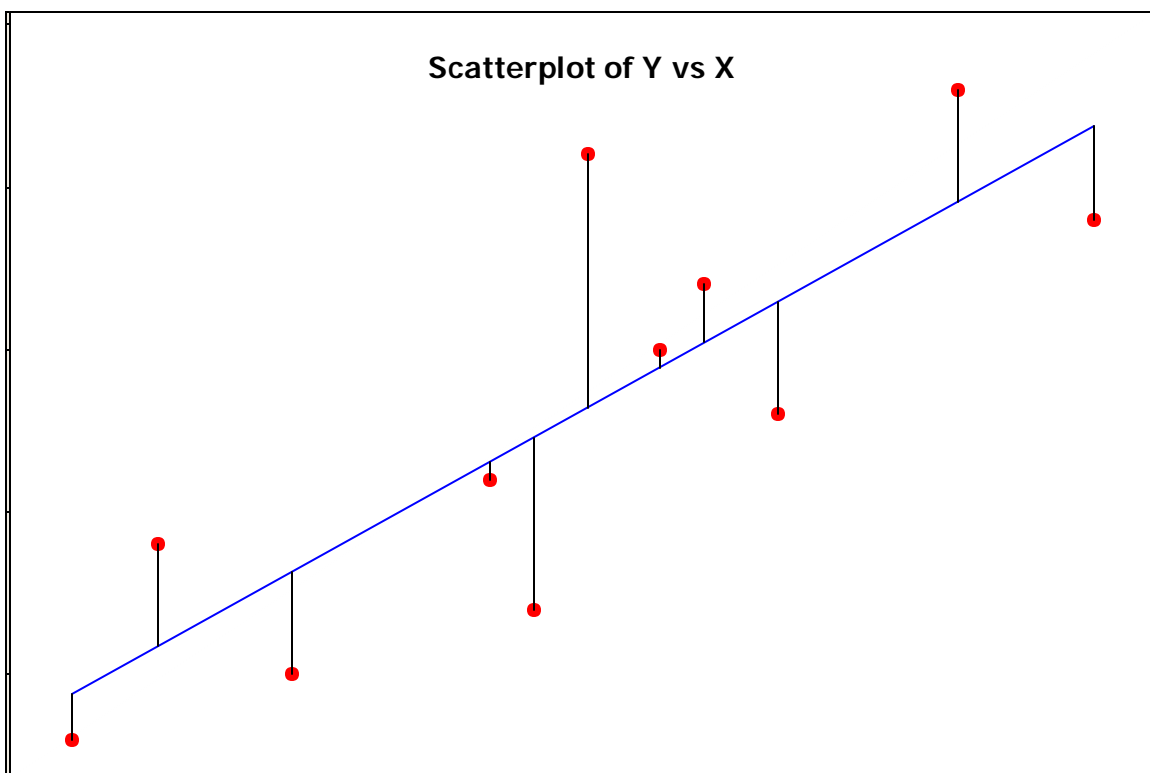
Uses and Misuses of R^2 and r in Curve Fitting

John L. Climent, Ph.D.
jcliment@cecilcc.edu

Before we begin any examples let's review the process of Least Squares Regression. For our Introductory Statistics students it is important for them to understand in concept what Least Squares Regression does (**it minimizes the sum of squared errors**) so that they can understand things that can affect both r and R^2 .



Given a random variable Y with fixed values of X , our regression equation is $\hat{Y} = b_0 + b_1 X$. We define our errors as $e_i = (y_i - \hat{y}_i) = (y_i - b_0 - b_1 x_i)$ and they are shown in the scatter plot below.



Now since $\sum e_i = \sum (y_i - \hat{y}_i) = 0$, we square each error and minimize $SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$.

This is just a Calculus 3 problem, that is of course over the heads of most of our students, but let's look briefly look at its solution.

For a given data set X and Y, find b_0 and b_1 so as to minimize

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2.$$

Differentiating SSE with respect to b_0 and b_1 , we get

$$\frac{\partial(SSE)}{\partial b_0} = -2 \sum (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial(SSE)}{\partial b_1} = -2 \sum (y_i - b_0 - b_1 x_i) x_i$$

Setting the partial derivatives equal to zero and rearranging the terms yields what is referred to in regression as the *normal equations*,

$$nb_0 + b_1 \sum x_i = \sum y_i$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i.$$

Solving for b_1 yields,

$$b_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}.$$

Now the regression equation passes through the point (\bar{x}, \bar{y}) , so one usually calculates b_0 from b_1 as

$$b_0 = \bar{y} - b_1 \bar{x}.$$

There are three things that one should look out for when fitting a curve to a data set:

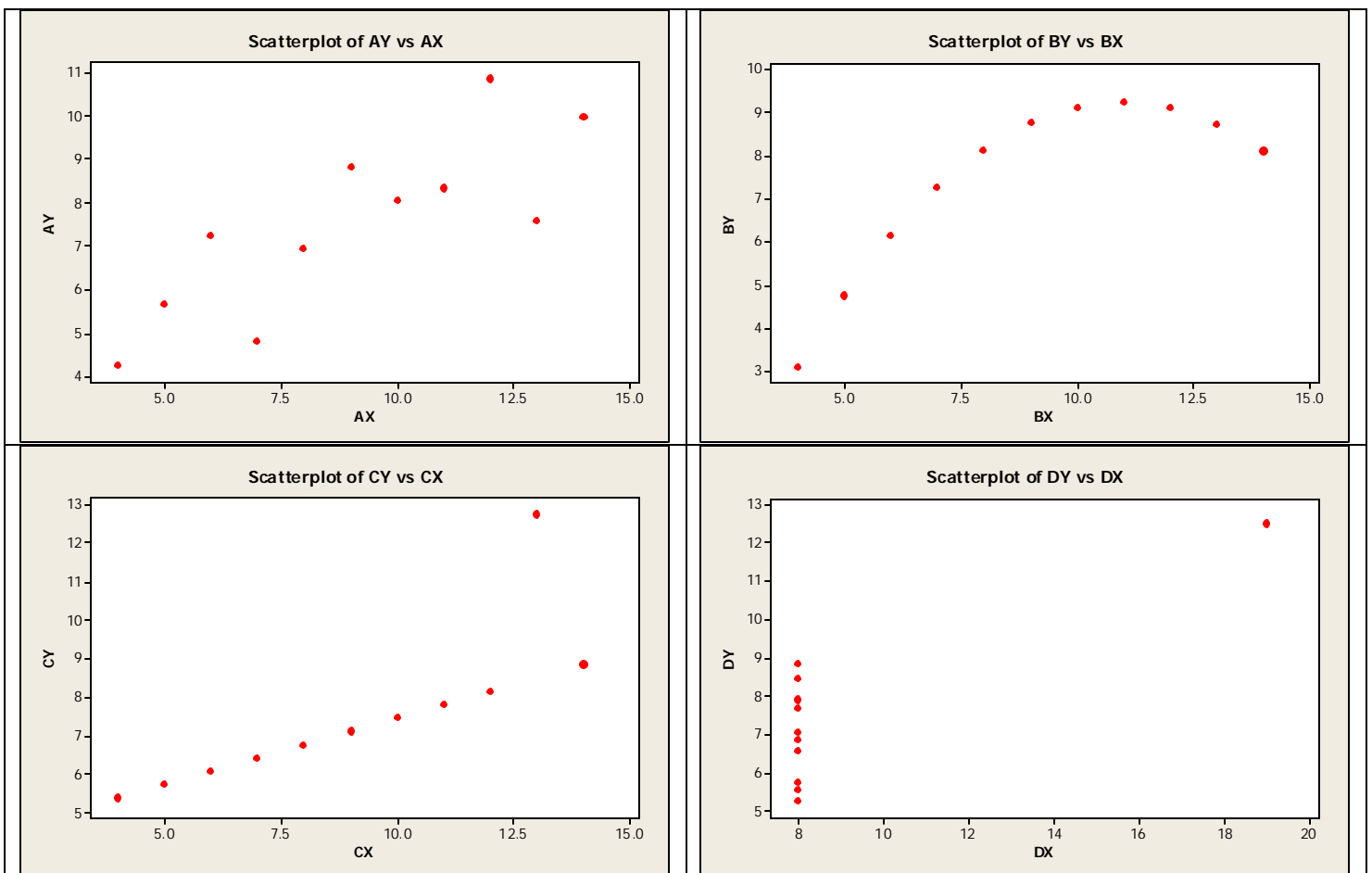
- **Using an inappropriate curve**
- **Outliers**
- **Influence points**

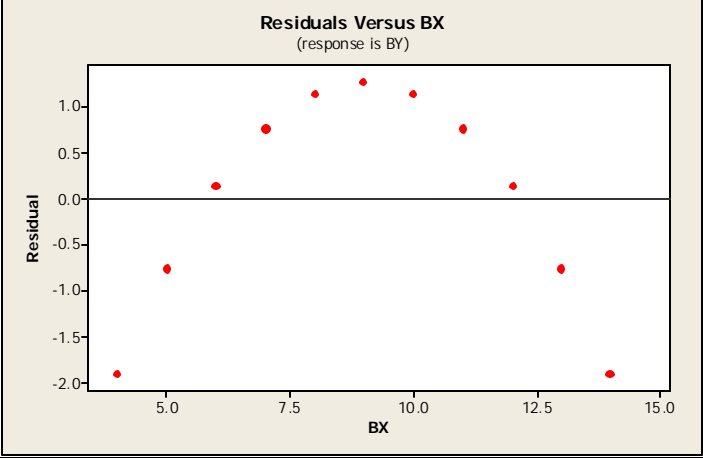
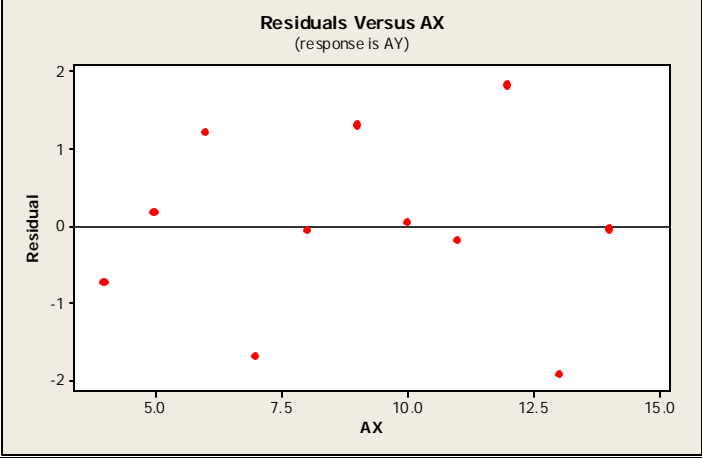
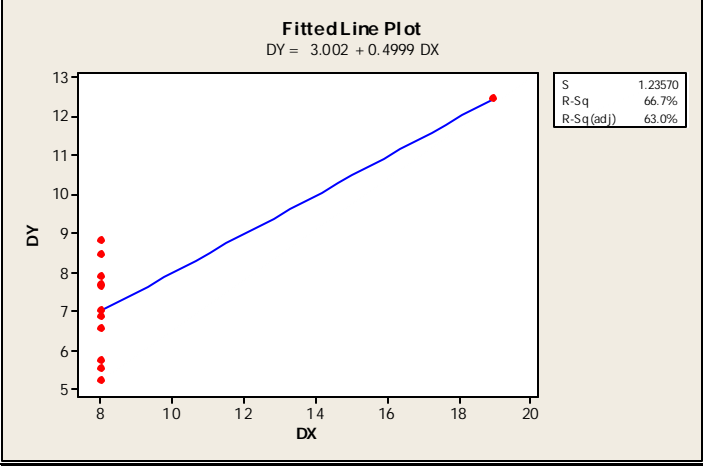
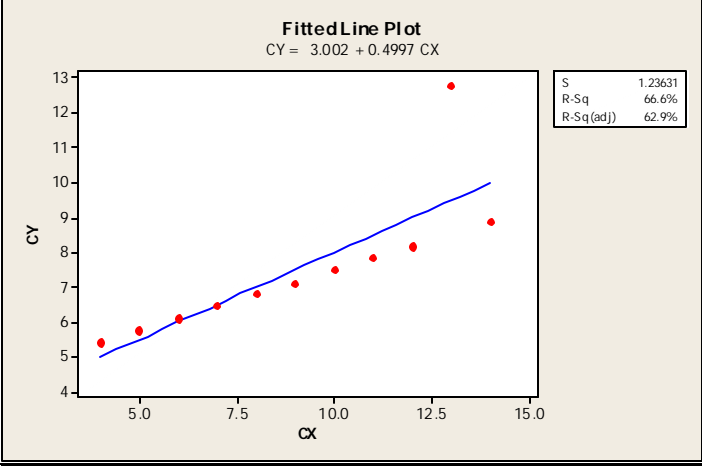
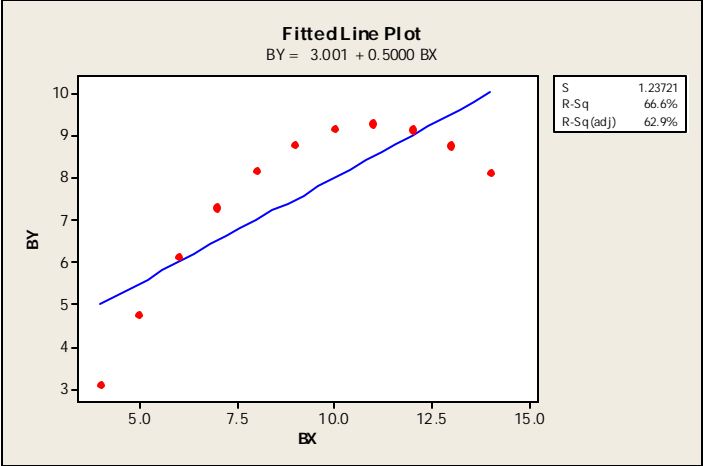
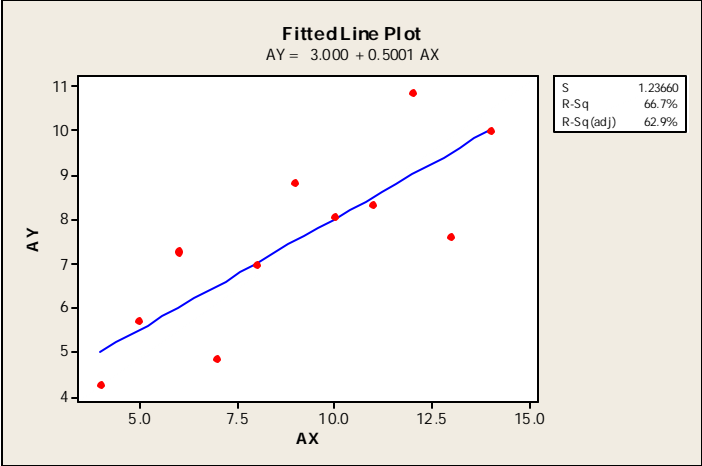
In 1973 F.J. Anscombe published a paper that illustrates the effect of all three of these things on regression. Anscombe published four data sets with identical numerical regression results, three of which were affected these items as follows:

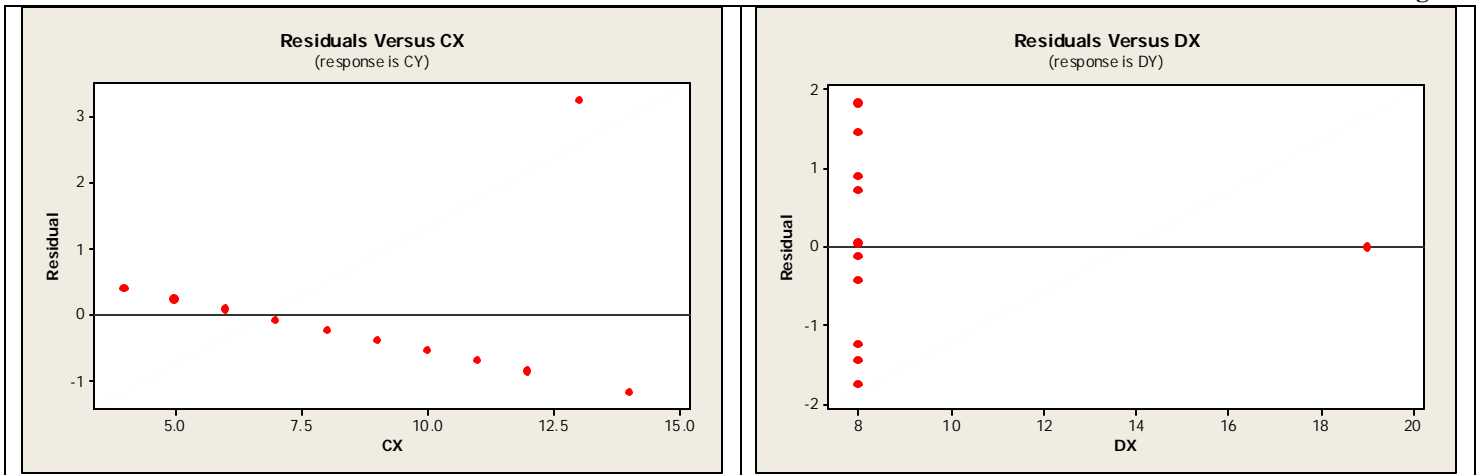
- Data Set A – an obvious linear model
- Data Set B – a curvilinear model
- Data Set C – affect of an outlier
- Data set D – the affect of an influence point

Data Set A		Data Set B		Data Set C		Data Set D	
AX	AY	BX	BY	CX	CY	DX	DY
10	8.04	10	9.14	10	7.46	8	6.58
14	9.96	14	8.10	14	8.84	8	5.76
5	5.68	5	4.74	5	5.73	8	7.71
8	6.95	8	8.14	8	6.77	8	8.84
9	8.81	9	8.77	9	7.11	8	8.47
12	10.84	12	9.13	12	8.15	8	7.04
4	4.26	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	7	6.42	19	12.50
11	8.33	11	9.26	11	7.81	8	5.56
13	7.58	13	8.74	13	12.74	8	7.91
6	7.24	6	6.13	6	6.08	8	6.89
$\hat{Y} = 3.0 + 0.5X$ $S_{YX} = 1.273$ $S_{b_1} = 0.118$ $r = 0.816$ $r^2 = 0.667$ $SSR = 27.51$ $SSE = 13.763$ $SST = 41.273$		$\hat{Y} = 3.0 + 0.5X$ $S_{YX} = 1.273$ $S_{b_1} = 0.118$ $r = 0.816$ $r^2 = 0.667$ $SSR = 27.51$ $SSE = 13.763$ $SST = 41.273$		$\hat{Y} = 3.0 + 0.5X$ $S_{YX} = 1.273$ $S_{b_1} = 0.118$ $r = 0.816$ $r^2 = 0.667$ $SSR = 27.51$ $SSE = 13.763$ $SST = 41.273$		$\hat{Y} = 3.0 + 0.5X$ $S_{YX} = 1.273$ $S_{b_1} = 0.118$ $r = 0.816$ $r^2 = 0.667$ $SSR = 27.51$ $SSE = 13.763$ $SST = 41.273$	

Source: F.J. Anscombe, "Graphs in Statistical Analysis" American Statistician 27 (1973); 17-21

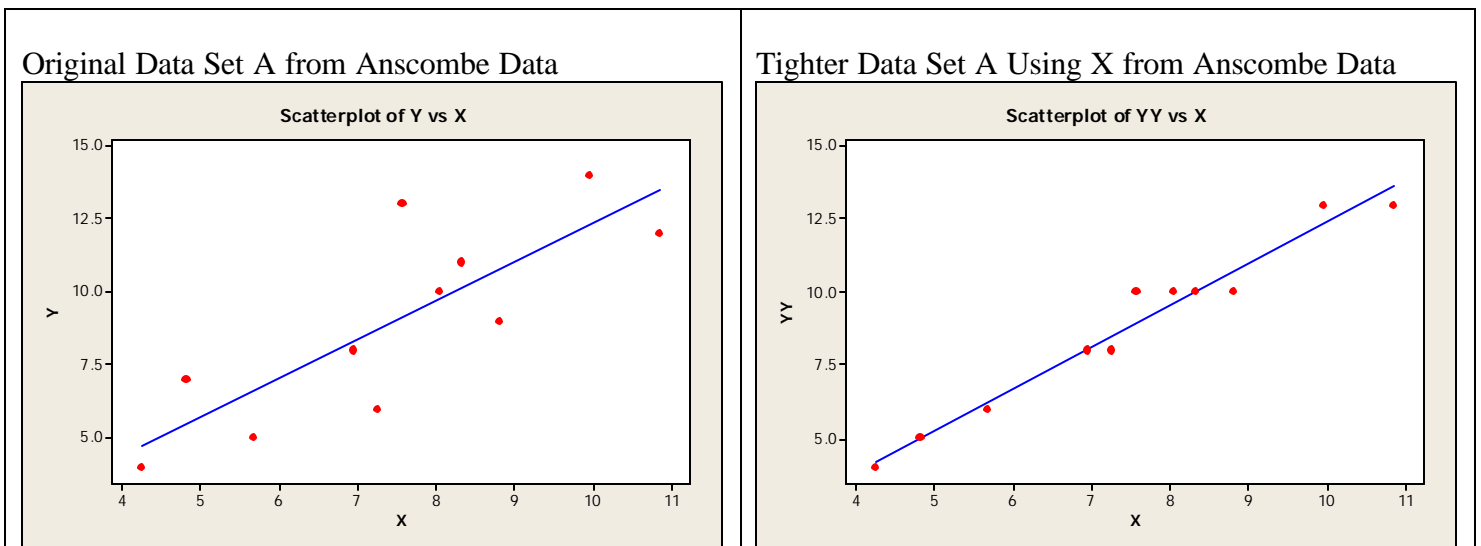




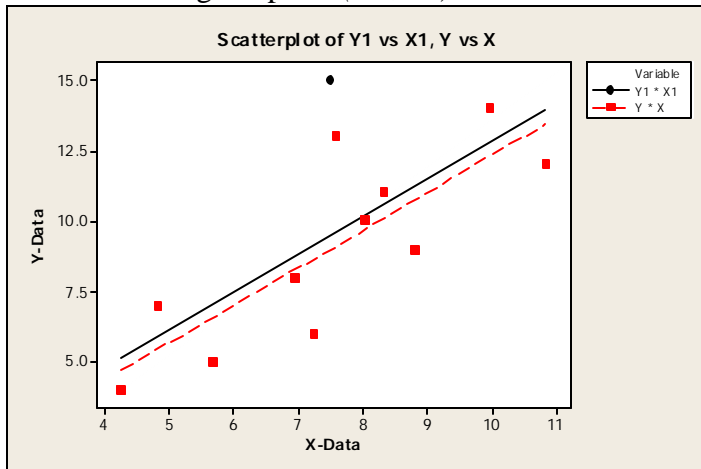


Let's create outliers and influence points. Data Set A from Anscombe will be used for this purpose; however, the X and Y were accidentally reversed when preparing the portion. Since the slides were already printed when this was discovered and since the role of X and Y are for the most part arbitrary here there will be no loss in using the data in this form to illustrate these principles. A tighter data set using the same X values was also used for comparison.

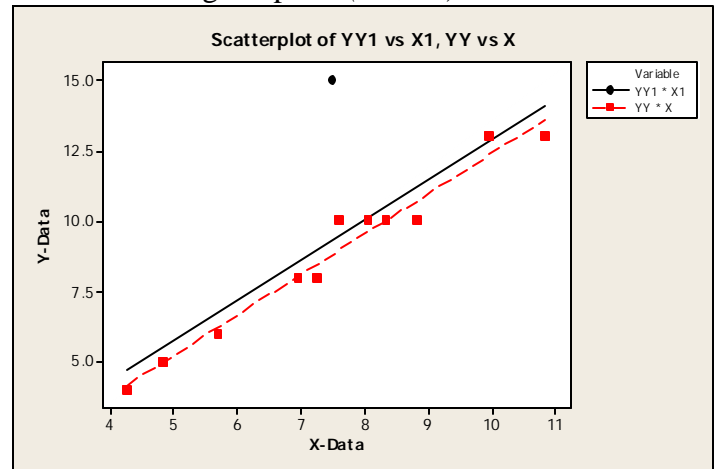
X	4.26	4.82	5.68	6.95	7.24	7.58	8.04	8.33	8.81	9.96	10.84
Y	4	7	5	8	6	13	10	11	9	14	12
YY	4	5	6	8	8	10	10	10	10	13	13



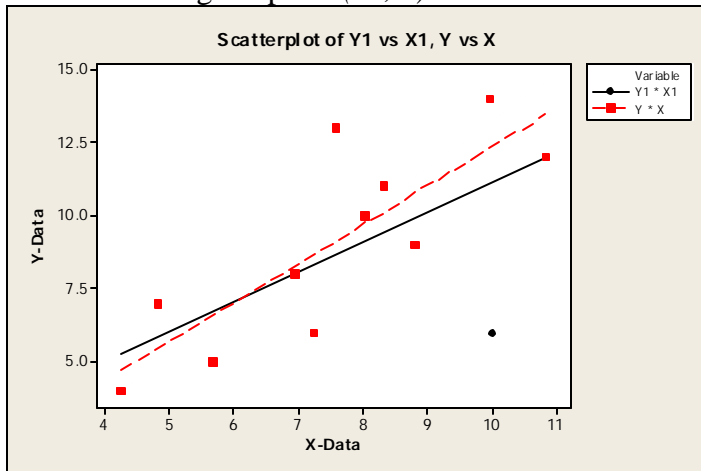
Effect of adding the point (7.5, 15) as an outlier.



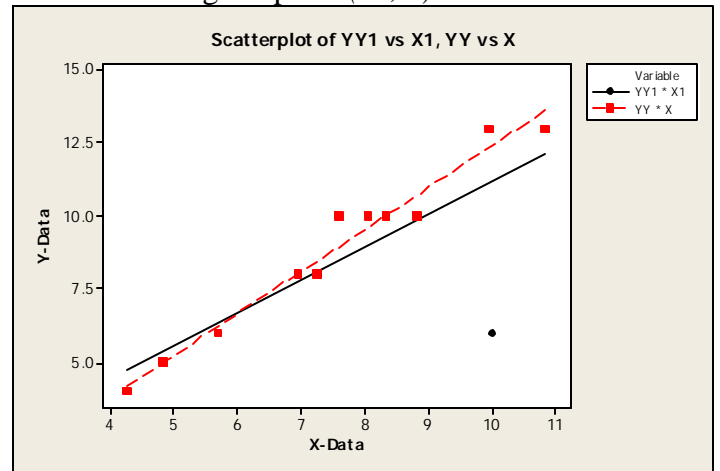
Effect of adding the point (7.5, 15) as an outlier.



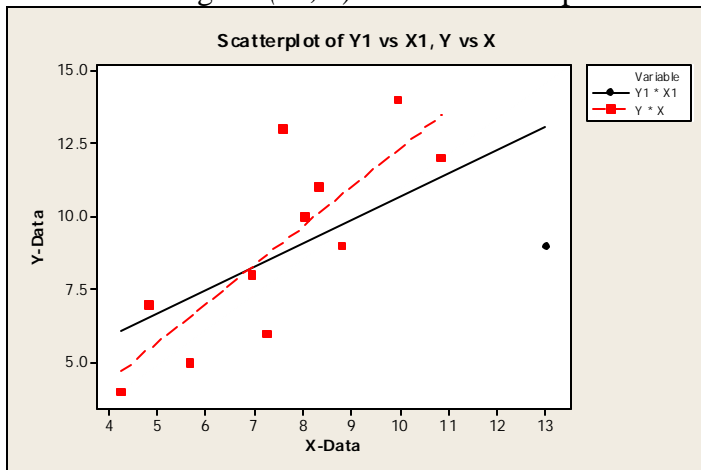
Effect of adding the point (10, 6) as an outlier.



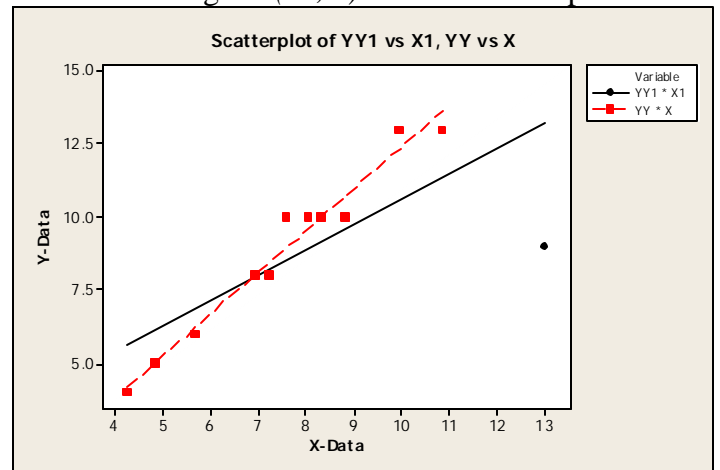
Effect of adding the point (10, 6) as an outlier.



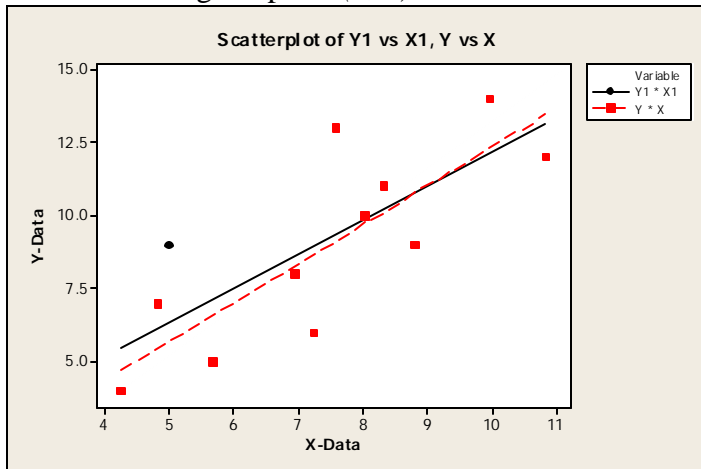
Effect of adding the (13, 9) as an influence point.



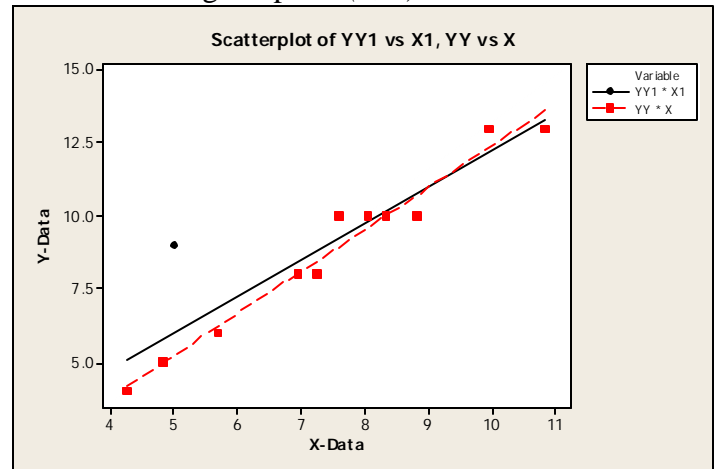
Effect of adding the (13, 9) as an influence point.



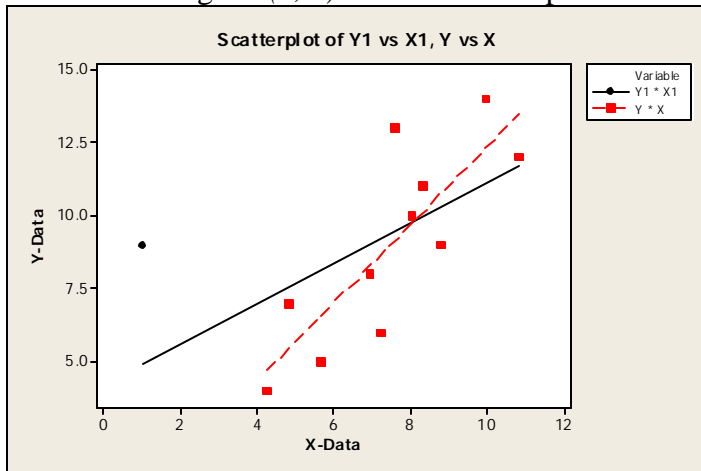
Effect of adding the point (5, 9) as an outlier.



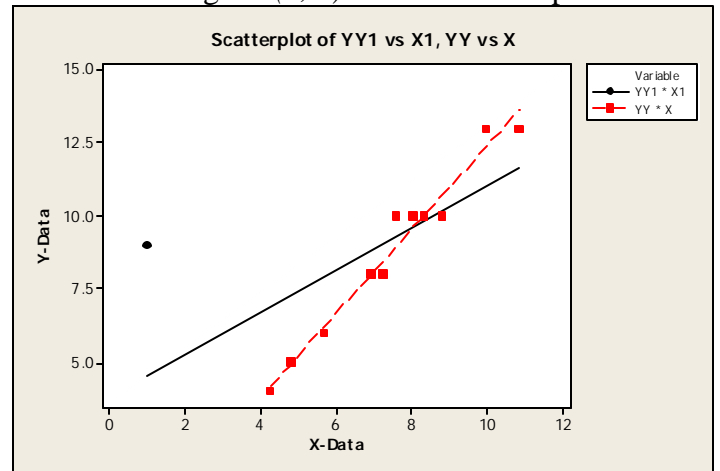
Effect of adding the point (5, 9) as an outlier.



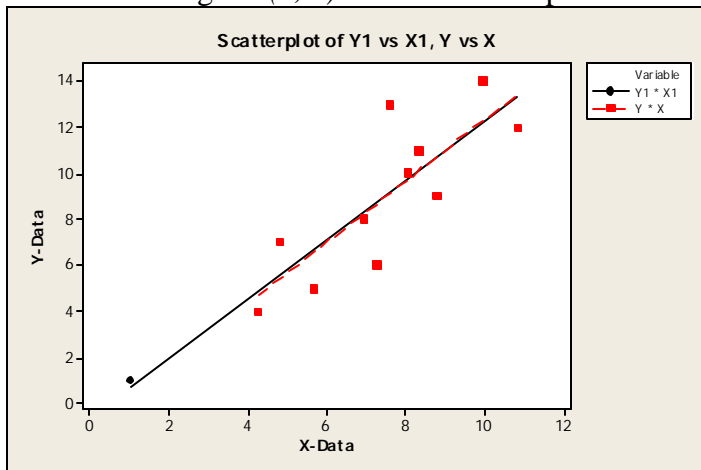
Effect of adding the (1, 9) as an influence point.



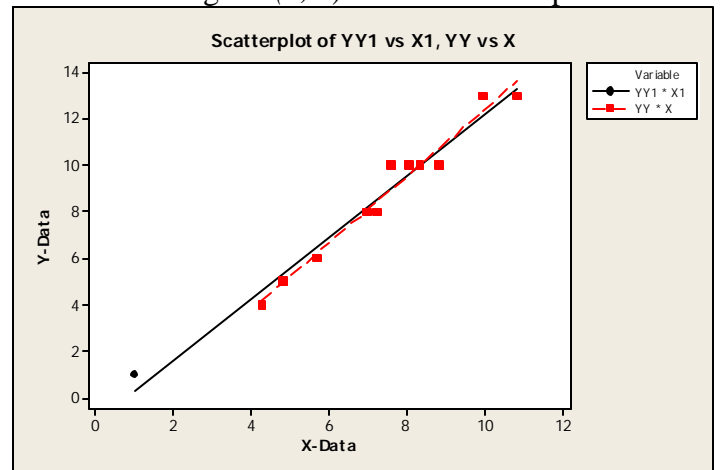
Effect of adding the (1, 9) as an influence point.



Effect of adding the (1, 1) as an influence point.



Effect of adding the (1, 1) as an influence point.



Principle 1: Students should never be asked to fit curves to abstract variables X and Y. Instead they should be given variables that they readily understand and they should be asked to fit the curve that makes the most sense for the variables presented.

Let's illustrate this with the age in years and asking price in thousands of dollars for Honda Civics advertised in a local newspaper.

Age	1	4	16	12	9	4	5	6	6	13	4	3	3	1	5	4
Price	15.4	8.8	2.7	2.2	4.3	10.8	8.8	5.0	8.7	1.6	13.0	9.5	12.9	14.4	10.5	11.5

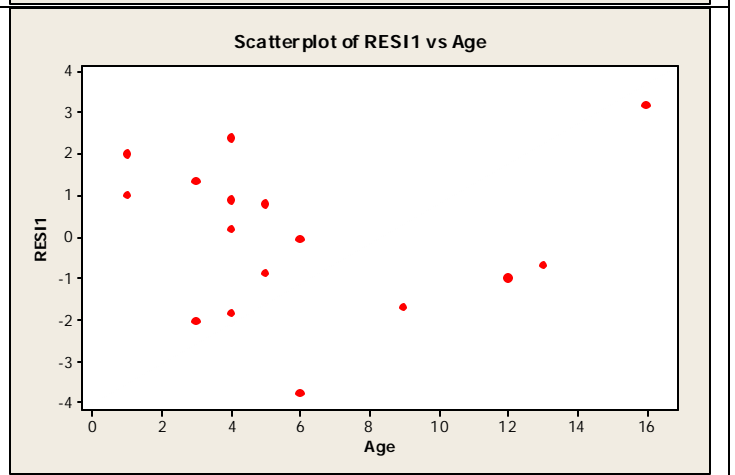
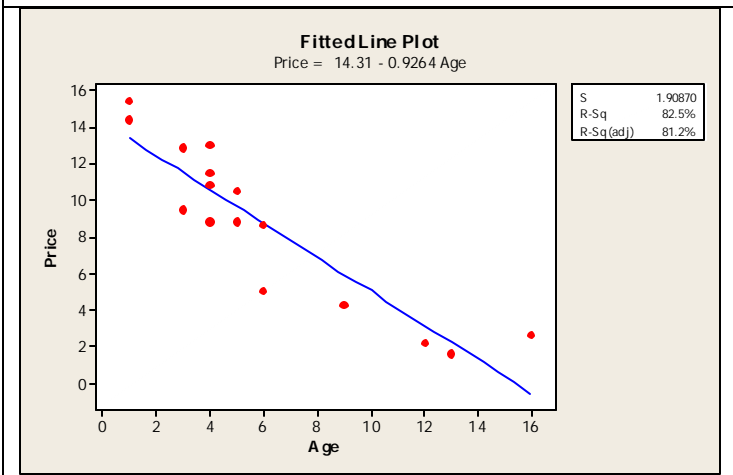
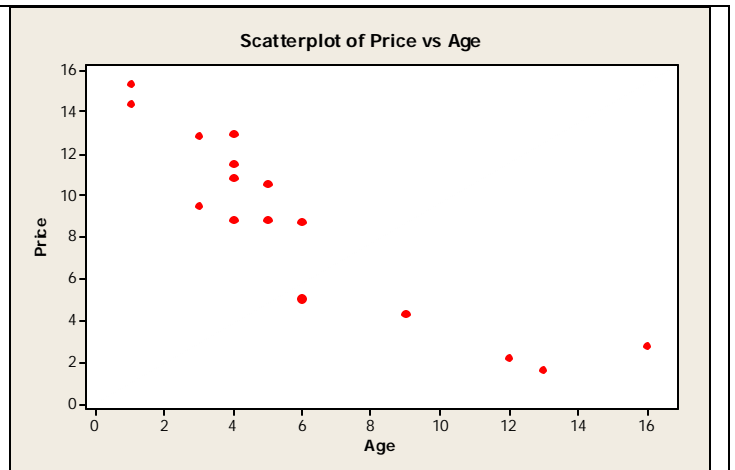
Linear Model
Regression Analysis: Price versus Age

The regression equation is
 Price = 14.31 - 0.9264 Age

S = 1.90870 R-Sq = 82.5% R-Sq(adj) = 81.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	240.316	240.316	65.96	0.000
Error	14	51.004	3.643		
Total	15	291.319			



Quadratic Model
Polynomial Regression Analysis: Price versus Age

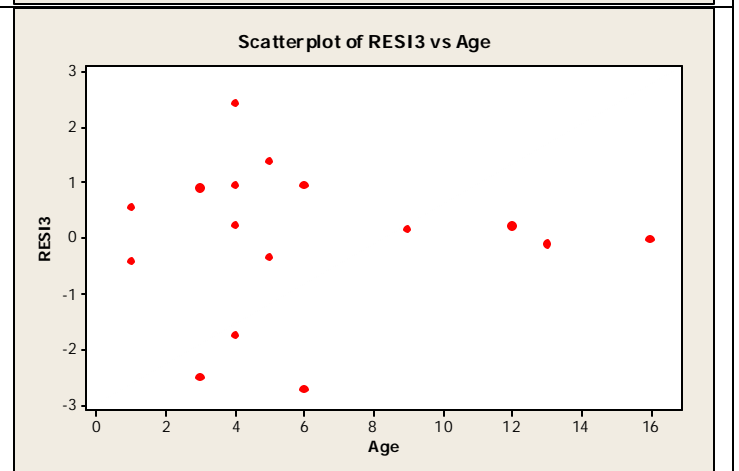
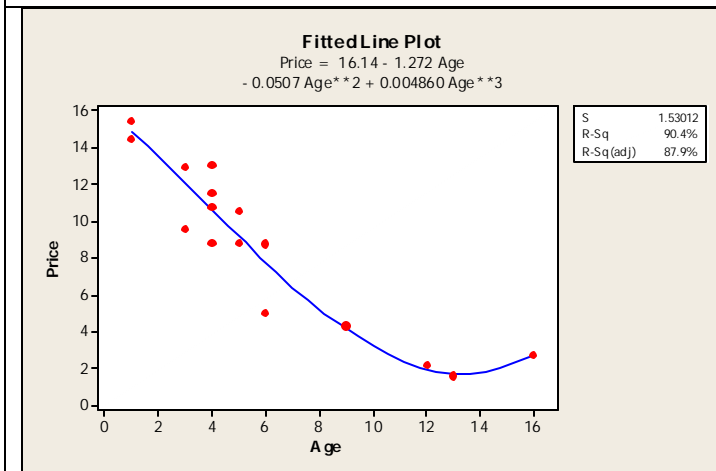
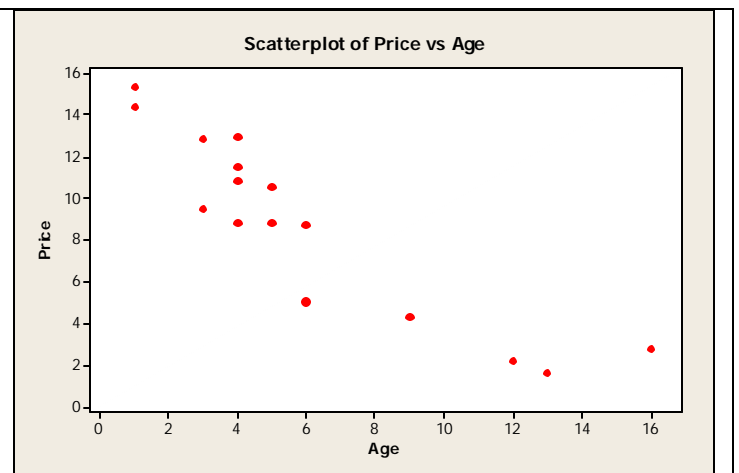
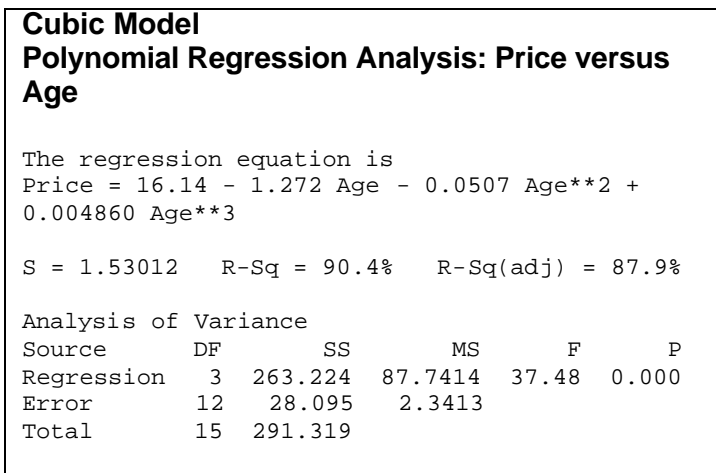
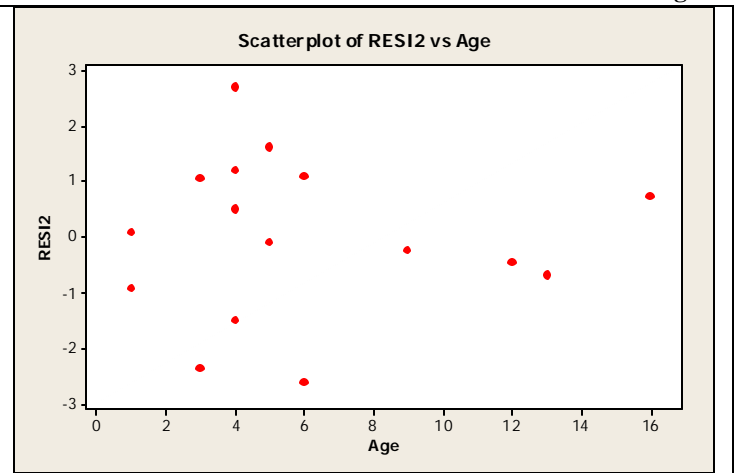
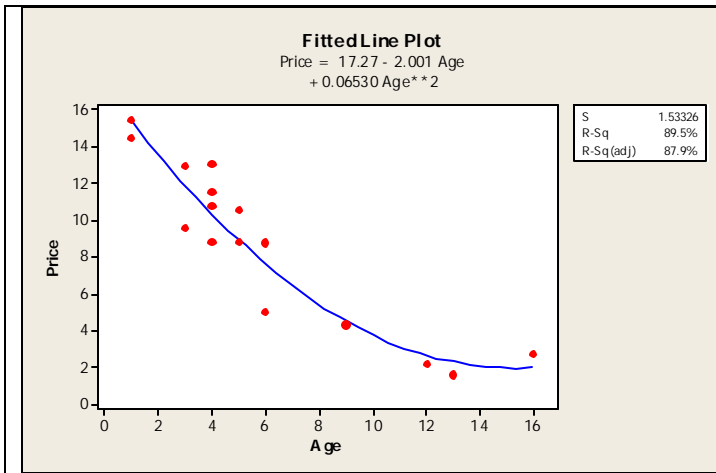
The regression equation is
 Price = 17.27 - 2.001 Age + 0.06530 Age**2

S = 1.53326 R-Sq = 89.5% R-Sq(adj) = 87.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	260.758	130.379	55.46	0.000
Error	13	30.562	2.351		
Total	15	291.319			





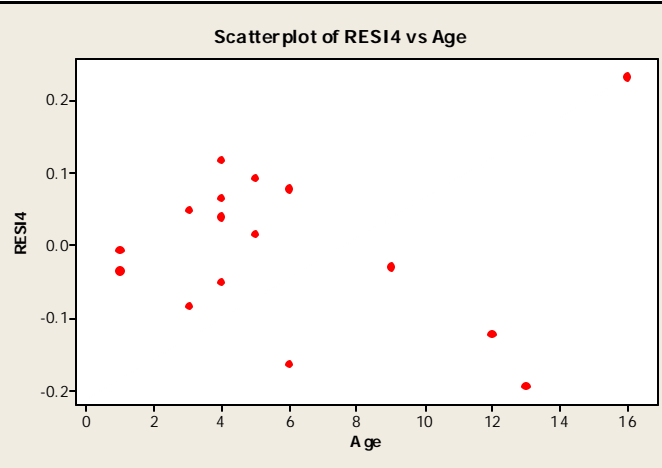
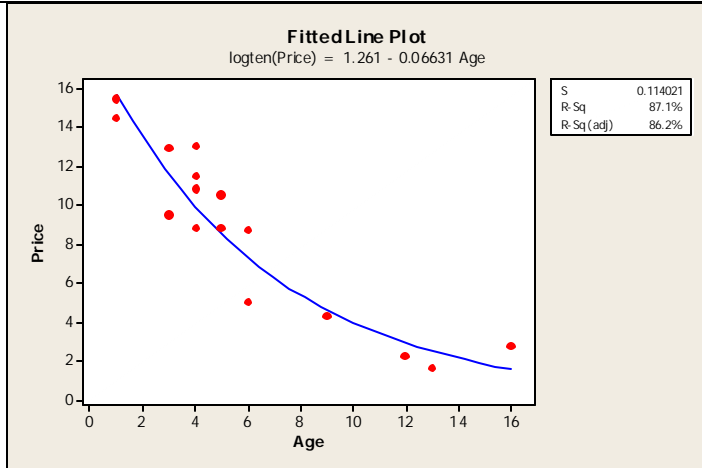
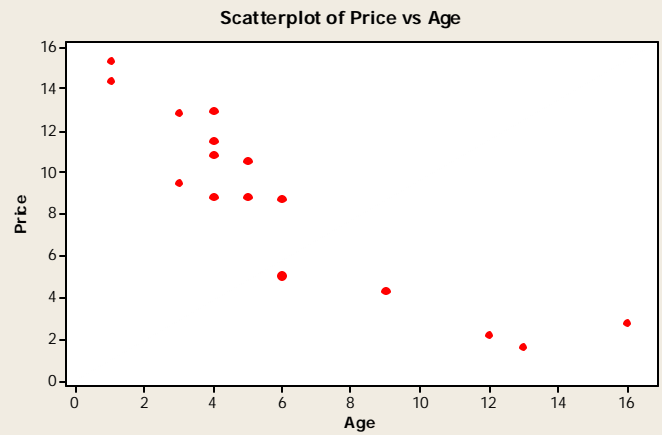
**Exponential Model
Regression Analysis: Price versus Age**

The regression equation is
 $\log_{10}(\text{Price}) = 1.261 - 0.06631 \text{ Age}$

S = 0.114021 R-Sq = 87.1% R-Sq(adj) = 86.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1.23128	1.23128	94.71	0.000
Error	14	0.18201	0.01300		
Total	15	1.41330			



Summary of Results for Price vs. Year

Model	Equation	R ²	R ² _{adj}	Ŷ(16)	Ŷ(20)
Linear	$\hat{Y} = 14.31 - 0.9264X$	82.5%	81.2%	-\$512	-\$4,218
Quadratic	$\hat{Y} = 17.27 - 2.001X + 0.06530X^2$	89.5%	87.9%	\$1,970	\$3,370
Cubic	$\hat{Y} = 16.14 - 1.272X - 0.0507X^2 + 0.004860X^3$	90.4%	87.9%	\$2,714	\$9,300
Exponential	$\log(\hat{Y}) = 1.261 - 0.06631X$	87.1%	86.2%	\$1,585	\$861

The exponential model is the best one here, because it explains the expected behavior of the variables. As the age of the Honda Civic increases the price will eventually level off and approach zero. By choosing the appropriate model, we are able to reasonably extrapolate to values of X outside the scope of our data.

An Explanation of R²:

Assume that our random variable is Y and that for the moment it is not associated with any values of X. We could simply use the variance of Y to measure its variability as follows.

$$VAR[Y] = s_Y^2 = \frac{\sum (y_i - \bar{y}_i)^2}{n-1}$$

Now for a given data set in X and Y , the number of data points, n , is fixed, so the numerator of the previous expression will serve just as well as a measure of the variability. The numerator is referred to as the total sum of squares (SST)

$$SST = \sum (y_i - \bar{y}_i)^2$$

Now let's introduce the variable X into the picture. Least squares regression will give us the equation $\hat{Y} = b_0 + b_1 X$. Note: \hat{Y} does not have to be a linear function here; it can be a polynomial function. We then add and subtract \hat{Y} in the above expression with the following results:

$$\begin{aligned} \sum (y_i - \bar{y}_i)^2 &= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y}_i)^2 \\ &= \sum [(y_i - \hat{y}_i) - (\bar{y}_i - \hat{y}_i)]^2 \\ &= \sum [(y_i - \hat{y}_i)^2 - 2(y_i - \hat{y}_i)(\bar{y}_i - \hat{y}_i) + (\bar{y}_i - \hat{y}_i)^2] \\ &= \sum (y_i - \hat{y}_i)^2 - 2 \sum (y_i - \hat{y}_i)(\bar{y}_i - \hat{y}_i) + \sum (\bar{y}_i - \hat{y}_i)^2 \end{aligned}$$

Now it turns out that with a lot of algebra the middle term of the above expression is zero, so we get

$$\sum (y_i - \bar{y}_i)^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\bar{y}_i - \hat{y}_i)^2$$

Defining SSE and SSR as the latter two terms in the above expression we get the following equivalent relationships.

$$\begin{aligned} SST &= SSE + SSR \\ 1 &= \frac{SSE}{SST} + \frac{SSR}{SST} \end{aligned}$$

We now define R^2 , the coefficient of determination as:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

Clearly $0=R^2=I$, and if $R^2=I$, $SSE=0$ then all of the points fall on the graph of \hat{Y} and \hat{Y} has explained all of the variability in the sample values. If $R^2=0$, $SSE=SST$ then none of the variability has been explained. In essence R^2 measures the portion or percent of the total variation explained by the regression equation \hat{Y} .

An Explanation of r :

Karl Pearson has been credited with developing the correlation coefficient, r . Generalizing on earlier work by Francis Galton, Pearson first wrote on the correlation coefficient in 1895. It has since become known as the Pearson's product-moment correlation coefficient or simply the correlation coefficient.

$$r = \frac{\sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Correlation analysis differs from regression analysis in a few fundamental ways. In regression analysis Y is considered our random variable, but X is considered to have fixed values. In correlation analysis both Y and X are considered to be random variables. The correlation coefficient, r , only measures the strength of the linear relationship between X and Y and it should not be used for nonlinear relationships. The coefficient of determination R^2 can be used for linear and nonlinear relationships. When R^2 is used for linear relationships then $R^2 = (r)^2$, but this relationship does not hold for nonlinear relationships. If one considers the population correlation coefficient ρ , as opposed to the sample correlation coefficient r , X and Y are considered to come from a bivariate normal distribution. In regression analysis only Y is assumed to be normally distributed since the values of X are assumed to be fixed. Actually Y only needs to be normal in order to find confidence intervals or perform hypothesis on the parameters. Assuming that X and Y have a bivariate normal distribution the correlation, ρ , between X and Y is defined as the covariance between X and Y divided by their standard deviations.

$$r = \frac{C(X, Y)}{s_X s_Y} = \frac{E[(X - m_X)(Y - m_Y)]}{s_X s_Y}$$

The Cauchy-Schwarz inequality states that

$$(E[XY])^2 = |E[XY]|^2 \leq E[X^2]E[Y^2]$$

As a corollary to this theorem we see that

$$|r_{X,Y}| \leq 1$$

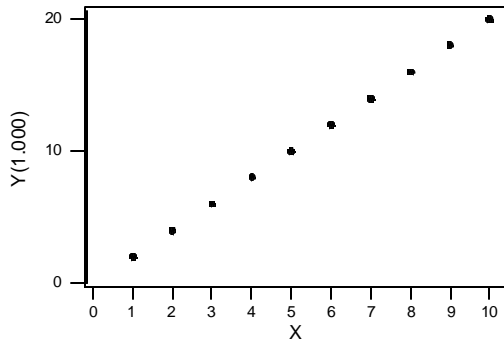
Showing students that linearity implies $r = \pm 1$:

Although showing that $r = \pm 1$ if and only if $y = b_0 + b_1x$ is too difficult to prove with our introductory students, it is not too difficult to show that if $y = b_0 + b_1x$ then $r = \pm 1$. Assume that $y = b_0 + b_1x$ then we have

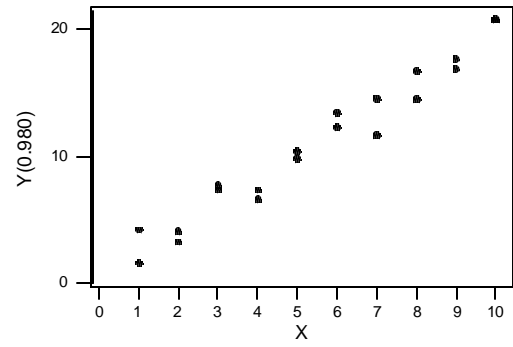
$$\begin{aligned} \bar{y} &= \frac{\sum y}{n} = \frac{\sum (b_0 + b_1x)}{n} = \frac{\sum b_0 + b_1 \sum x}{n} = \frac{nb_0 + b_1 \sum x}{n} = b_0 + b_1\bar{x} \\ y - \bar{y} &= (b_0 + b_1x) - (b_0 + b_1\bar{x}) = b_1(x - \bar{x}) \\ r &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}} = \frac{\sum (x - \bar{x}) b_1(x - \bar{x})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum [b_1(x - \bar{x})]^2}} \\ r &= \frac{b_1 \sum (x - \bar{x})^2}{\sqrt{\sum (x - \bar{x})^2} \sqrt{b_1^2 \sum (x - \bar{x})^2}} = \frac{b_1 \sum (x - \bar{x})^2}{|b_1| \left[\sqrt{\sum (x - \bar{x})^2} \right]^2} \\ r &= \frac{b_1 \sum (x - \bar{x})^2}{|b_1| \sum (x - \bar{x})^2} = \frac{b_1}{|b_1|} \\ r &= \frac{b_1}{|b_1|} = \begin{cases} 1, & \text{if } b_1 > 0 \\ -1, & \text{if } b_1 < 0 \end{cases} \end{aligned}$$

We know that r near zero indicates no linear relationship, but just how close to zero must we get? The following scatter plots using a positive slope can give us a good indication.

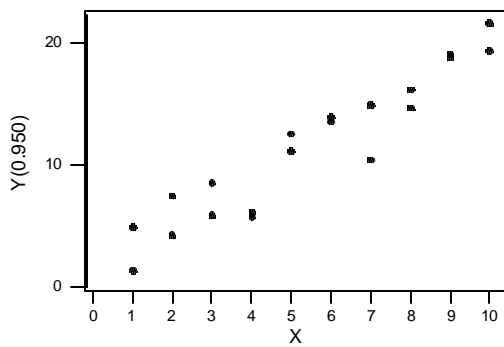
Correlation Coefficient $r = 1.000$



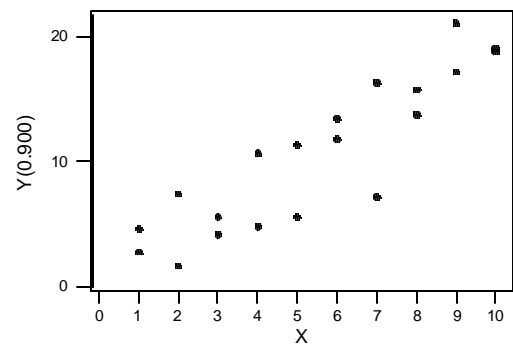
Correlation Coefficient $r = 0.980$



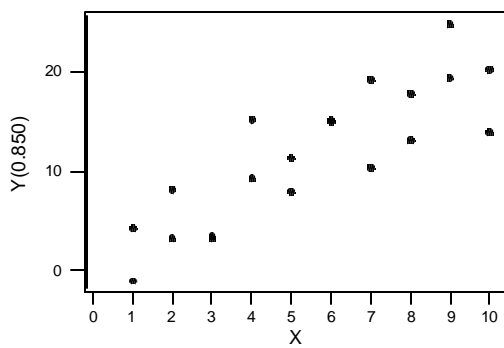
Correlation Coefficient $r = 0.950$



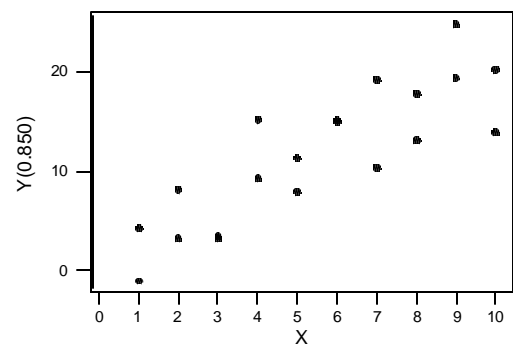
Correlation Coefficient $r = 0.900$



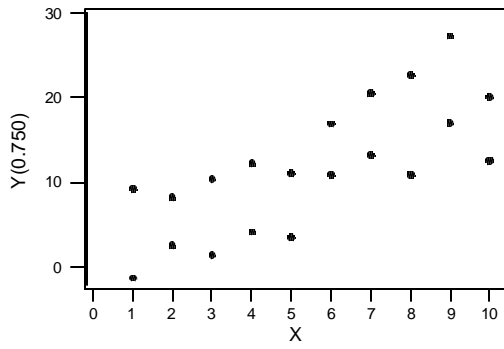
Correlation Coefficient $r = 0.850$



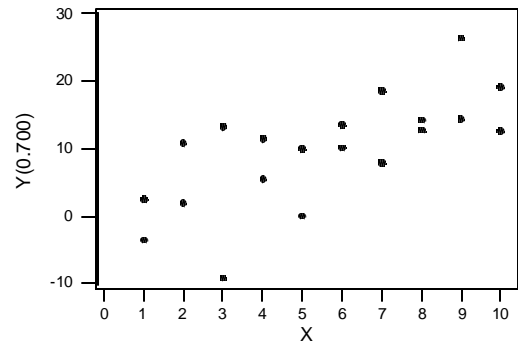
Correlation Coefficient $r = 0.800$



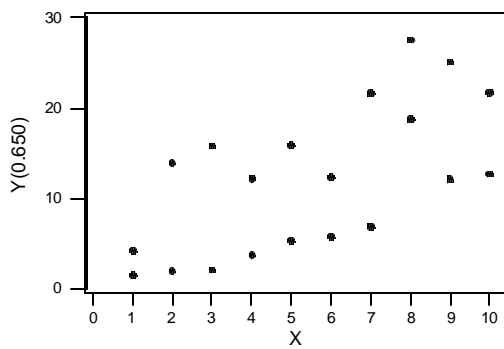
Correlation Coefficient r = 0.750



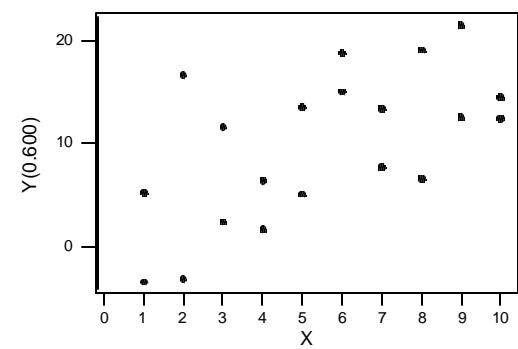
Correlation Coefficient r = 0.700



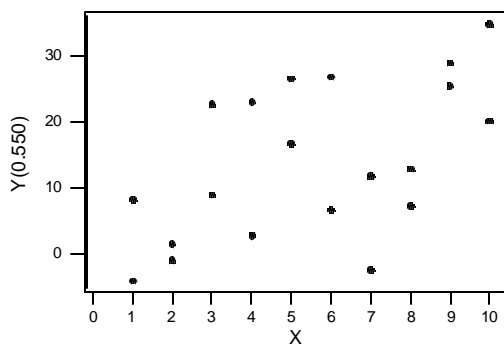
Correlation Coefficient r = 0.650



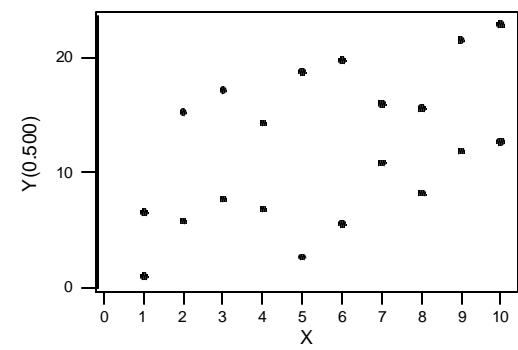
Correlation Coefficient r = 0.600



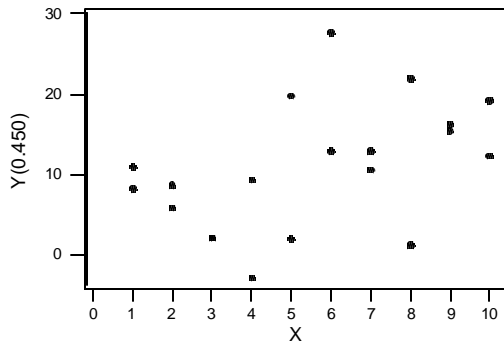
Correlation Coefficient r = 0.550



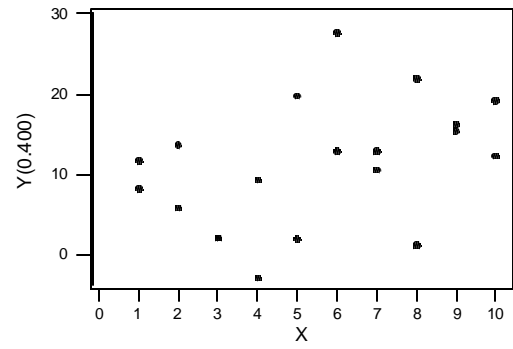
Correlation Coefficient r = 0.500



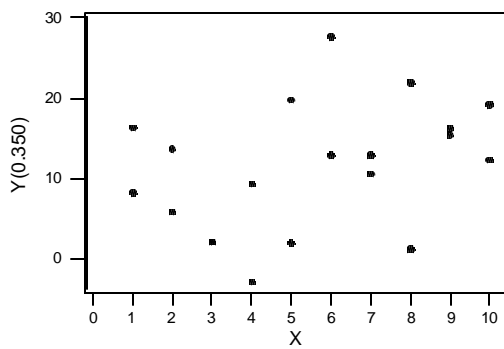
Correlation Coefficient r = 0.450



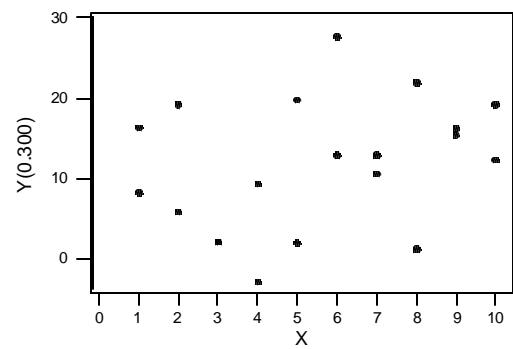
Correlation Coefficient r = 0.400



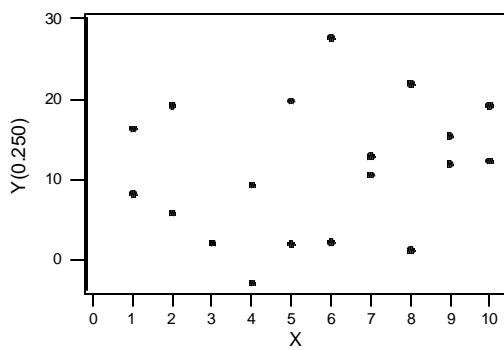
Correlation Coefficient r = 0.350



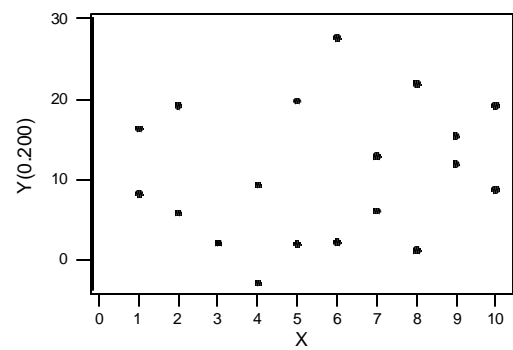
Correlation Coefficient r = 0.300



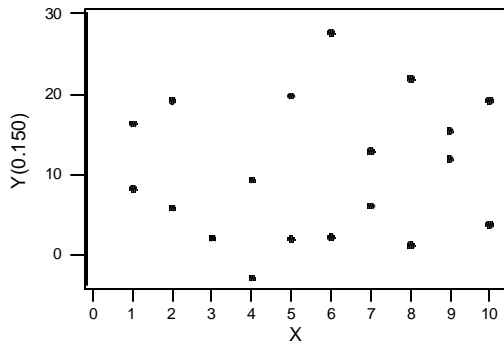
Correlation Coefficient r = 0.250



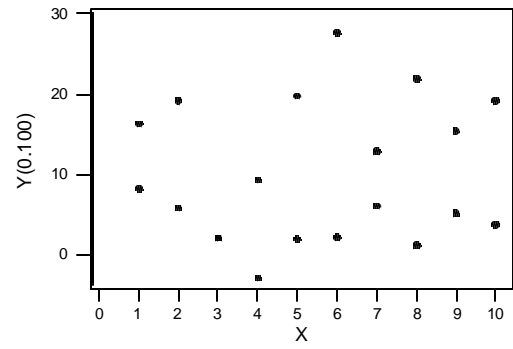
Correlation Coefficient r = 0.200



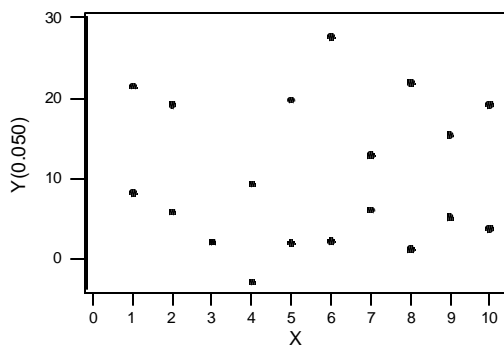
Correlation Coefficient $r = 0.150$



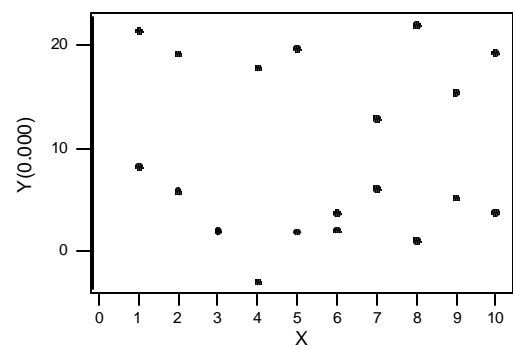
Correlation Coefficient $r = 0.100$



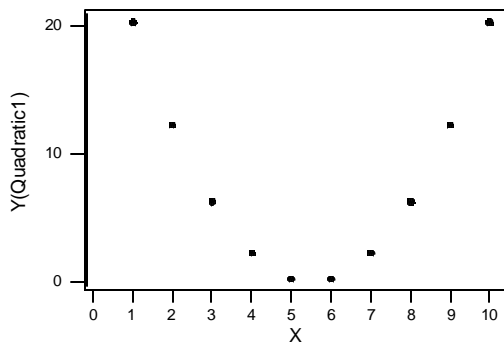
Correlation Coefficient $r = 0.050$



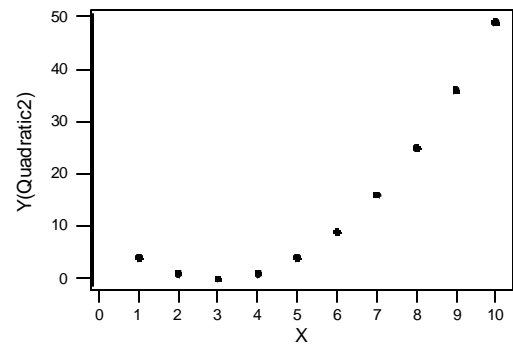
Correlation Coefficient $r = 0.000$



Curvilinear Relationship - Inappropriate to use r
(Calculated Correlation Coefficient $r = 0.000$)



Curvilinear Relationship - Inappropriate to use r
(Calculated Correlation Coefficient $r = 0.892$)

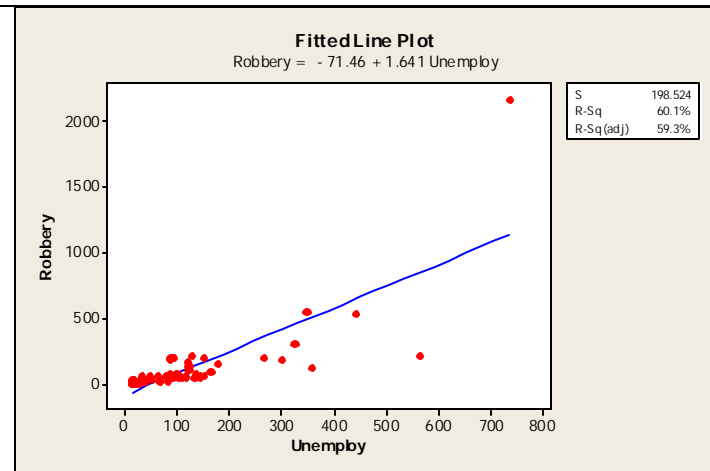


High correlation does not necessarily indicate cause and effect:

Let's look at the following data set to see this.

Bank Robbery Data				
Number	State	Robbery	Unemployed	Population
1	CT	59	64	3239
2	ME	17	25	1222
3	MA	208	127	5913
4	NH	5	21	1107
5	RI	14	21	998
6	VT	3	11	567
7	NJ	90	163	7736
8	NY	534	442	17950
9	PA	198	264	12040
10	IL	117	359	11658
11	IN	85	136	5593
12	MI	300	326	9273
13	OH	187	300	10907
14	WI	58	114	4867
15	IA	17	65	2840
16	KS	32	52	2513
17	MN	53	102	4353
18	MO	58	143	5159
19	NE	20	25	1611
20	ND	0	14	660
21	SD	3	15	715
22	DE	28	13	673
23	DC	29	16	604
24	FL	544	348	12671
25	GA	154	177	6436
26	MD	204	93	4694
27	NC	171	119	6571
28	SC	62	80	3512
29	VA	112	123	6098
30	WV	23	66	1857
31	AL	47	134	4118
32	KY	44	108	3727
33	MS	45	91	2621
34	TN	135	121	4940
35	AR	29	82	2406
36	LA	65	151	4382
37	OK	77	85	3224
38	TX	208	567	16991
39	AZ	198	89	3556
40	CO	76	98	3317
41	ID	7	25	1014
42	MT	13	24	806
43	NV	57	30	1111
44	NM	68	46	1528
45	UT	29	37	1707
46	WY	1	15	475
47	AK	5	17	527
48	CA	2161	737	29063
49	HI	30	13	1112
50	OR	190	84	2820
51	WA	199	151	4761

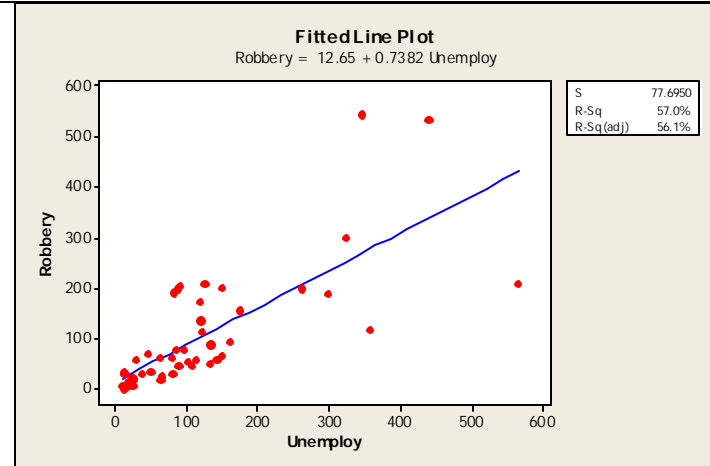
Bank Robbery Data - Rates of Robbery & Unemployment	
Robbery Rate	Unemployment Rate
0.0182155	0.0197592
0.0139116	0.0204583
0.0351767	0.0214781
0.0045167	0.0189702
0.0140281	0.0210421
0.0052910	0.0194004
0.0116339	0.0210703
0.0297493	0.0246240
0.0164452	0.0219269
0.0100360	0.0307943
0.0151976	0.0243161
0.0323520	0.0351558
0.0171450	0.0275053
0.0119170	0.0234231
0.0059859	0.0228873
0.0127338	0.0206924
0.0121755	0.0234321
0.0112425	0.0277186
0.0124146	0.0155183
0.0000000	0.0212121
0.0041958	0.0209790
0.0416048	0.0193165
0.0480132	0.0264901
0.0429327	0.0274643
0.0239279	0.0275016
0.0434597	0.0198125
0.0260234	0.0181099
0.0176538	0.0227790
0.0183667	0.0201705
0.0123856	0.0355412
0.0114133	0.0325401
0.0118057	0.0289777
0.0171690	0.0347196
0.0273279	0.0244939
0.0120532	0.0340815
0.0148334	0.0344592
0.0238834	0.0263648
0.0122418	0.0333706
0.0556805	0.0250281
0.0229123	0.0295448
0.0069034	0.0246548
0.0161290	0.0297767
0.0513051	0.0270027
0.0445026	0.0301047
0.0169889	0.0216755
0.0021053	0.0315789
0.0094877	0.0322581
0.0743557	0.0253587
0.0269784	0.0116906
0.0673759	0.0297872
0.0417979	0.0317160



Observe that one state, California, is an influence point an outlier. We see that the correlation is

Correlations: Robbery, Unemploy
 Pearson correlation of Robbery and Unemploy = 0.776
 P-Value = 0.000

Let's delete California from this data set.



For the data set without California, we see that the correlation is

Correlations: Robbery, Unemploy
 Pearson correlation of Robbery and Unemploy = 0.755
 P-Value = 0.000

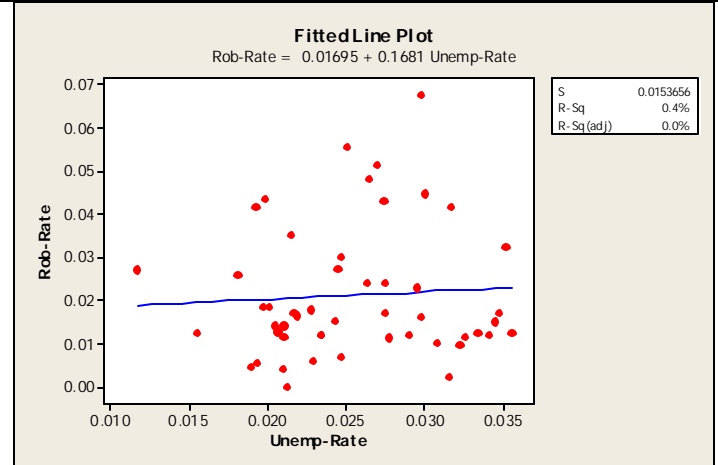
One might guess that Population is a lurking or confounding variable here. We can see this through the correlations between these three variables.

Correlations: Robbery, Unemploy, Pop

	Robbery	Unemploy
Unemploy	0.755	
Pop	0.804	0.971

Cell Contents: Pearson correlation

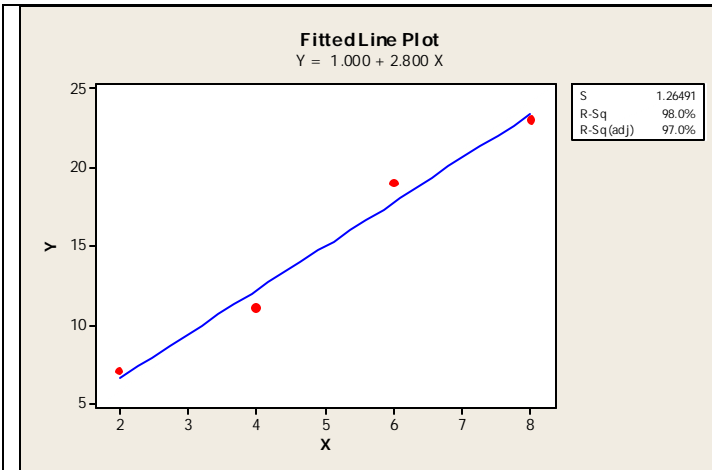
Actually we should be using the Unemployment Rate and the Bank Robbery Rate for our two variables. When we do this we see that there is no relationship between these two rates; thus, there was no cause and effect from our original two variables.



One Last Example:

Principle 2: Simply adding more terms to the regression equation tends to “artificially” increase the value of R^2 .

X	2	4	6	8
Y	7	11	19	23

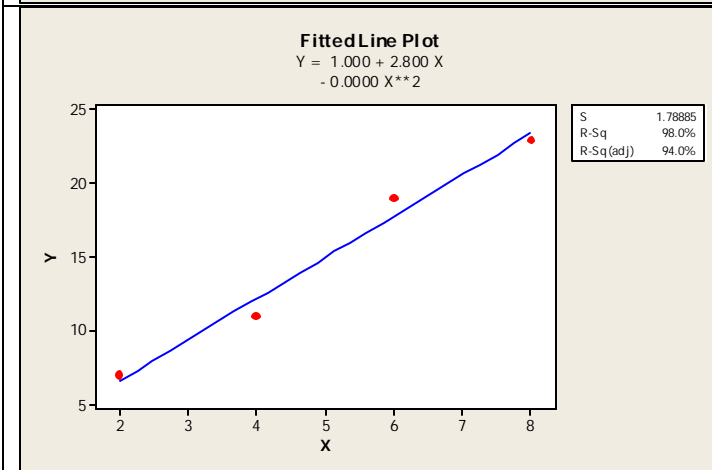


Regression Analysis: Y versus X
The regression equation is
Y = 1.000 + 2.800 X

S = 1.26491 R-Sq = 98.0% R-Sq(adj) = 97.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	156.8	156.8	98.00	0.010
Error	2	3.2	1.6		
Total	3	160.0			



Polynomial Regression Analysis: Y versus X
The regression equation is
Y = 1.000 + 2.800 X - 0.0000 X**2

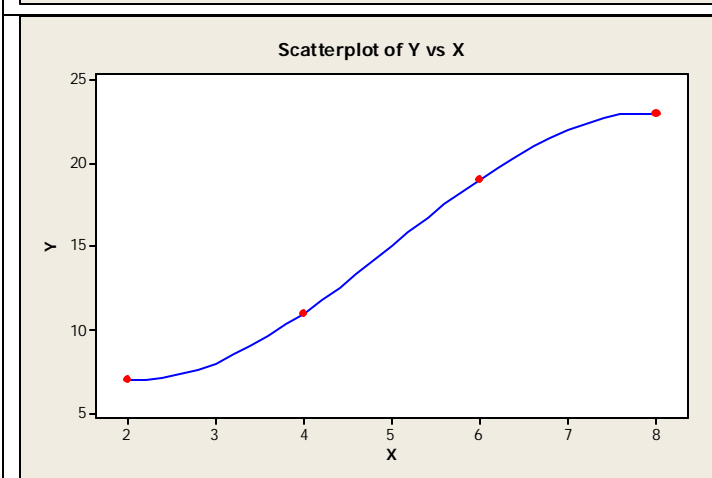
S = 1.78885 R-Sq = 98.0% R-Sq(adj) = 94.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	156.8	78.4	24.50	0.141
Error	1	3.2	3.2		
Total	3	160.0			

Sequential Analysis of Variance

Source	DF	SS	F	P
Linear	1	156.8	98.00	0.010
Quadratic	1	0.0	0.00	1.000



Note: Here $R^2 = 1$, but Minitab is programmed to not this model with too few values.