

Pitfalls of Regression

Tab 9:

Tab 9: Pitfalls of Regression

PURPOSE:

While Regression is a powerful tool, it must be used with care. This section will discuss common pitfalls and shortcomings of regression.

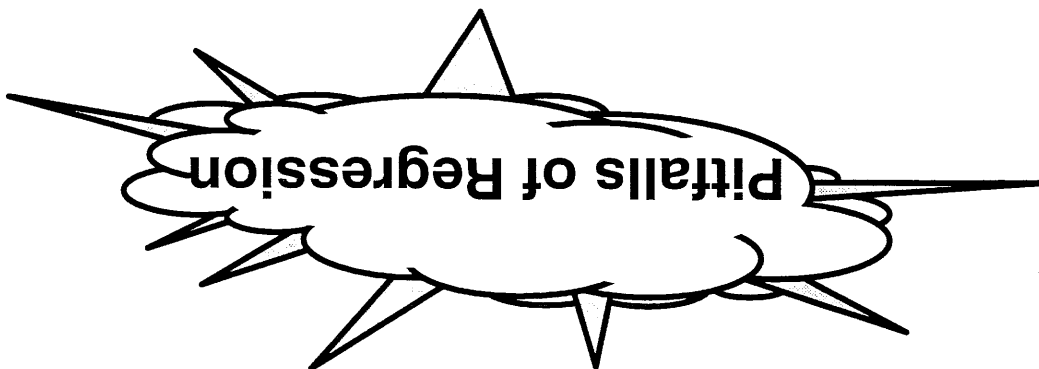
OBJECTIVES:

- Recognize situations where regression is used incorrectly
- Determine how to overcome the pitfalls of regression

WARNING!! BEWARE THE PITFALLS OF REGRESSION



PIT



Regression is a powerful tool, but it is often used incorrectly.

1. Correlation does not mean causation

2. Fitting the wrong model form

3. Relationships between independent variables (multi-collinearity)

4. Over-fitting; multiple hypothesis tests; too many independent variables

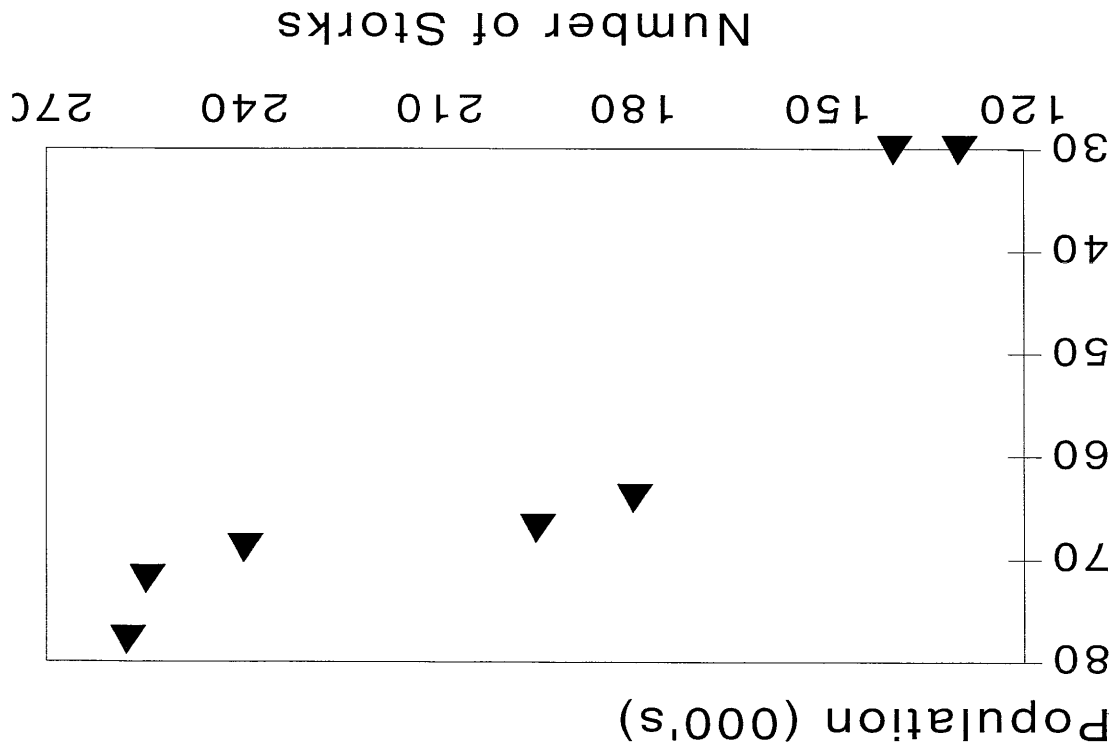
5. Influence of a few extreme values

6. Drawing firm conclusions from passive / happenstance data

7. Regression is used to prove statistically what you see from graphs - ALWAYS graph your data first!

Avoid the pitfalls of regression...always graph the data

1. Correlation does not mean causation



Even though the correlation coefficient (r) for the data is 0.918, killing storks would not be a good means of birth control.

Controlling the "X" will not affect the "Y"

2. Fitting the wrong model form

(These data are from Don Olsson)

Data:	
\bar{X}	\bar{Y}
10	30.5
20	16.8
50	7.9
100	4.8

Straight line

$$y = 25.8 - 0.241 X$$

$$r = 0.85$$

Quadratic

$$y = 36.1 - 0.881 X + 0.0057X^2$$

minimum at $x = 77$

$$r = 0.97$$

Cubic

$$Y = 52.3 - 2.64 X + 0.0484 X^2 - 0.000268 X^3$$

maximum at $x = 79$

$$r = 1.0$$

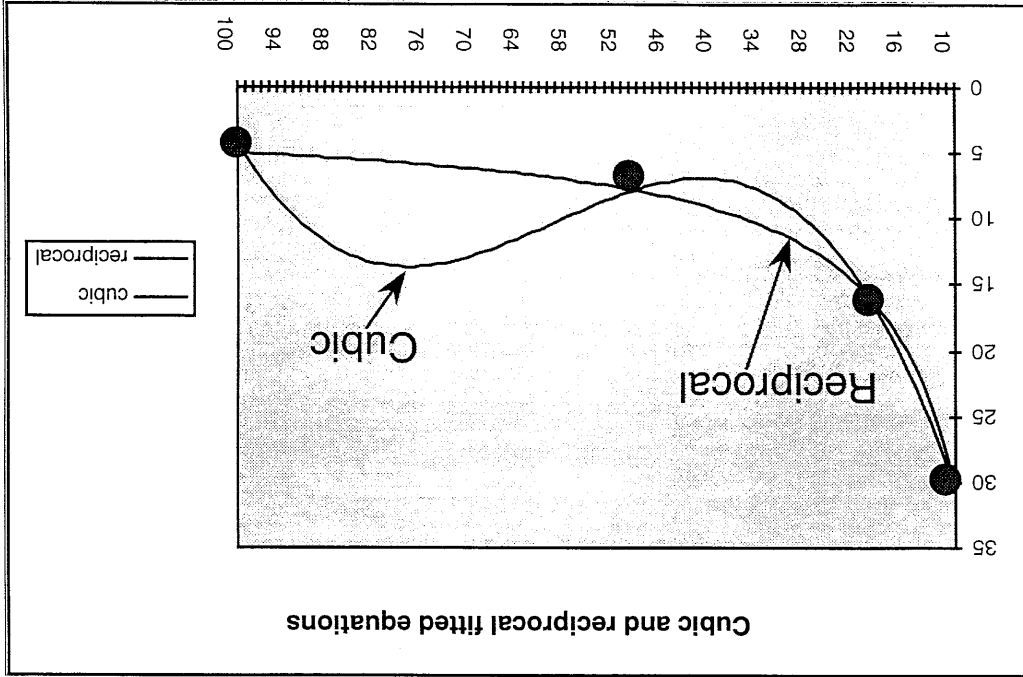
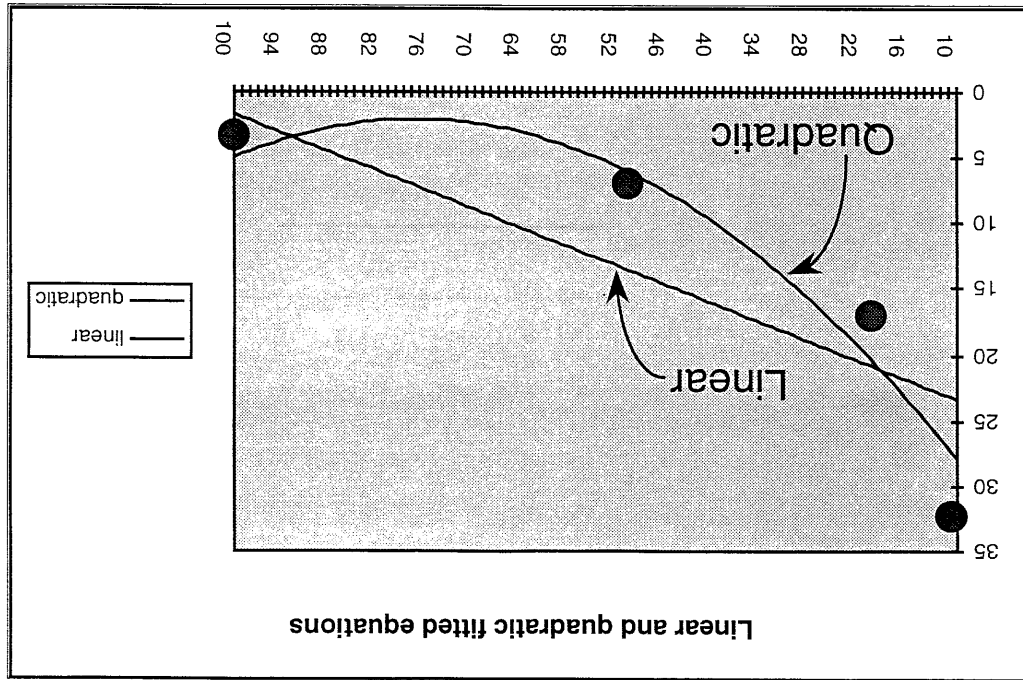
Correct model

$$y = 2.17 + 285 (1/x)$$

$$r = 0.9997$$

This "correct" model has the advantages of:

- Consistent with theory
- Provides an excellent fit
- Few parameters, simple form
- Model parameters have a physical meaning
- Interpolation is more likely to be valid




Remember that we are modeling with only 4 sets of data points!

3. Relationships between independent variables (multi-collinearity)

Percent ontime for a dishwasher cycle extender was measured at six different combinations of voltage and temperature.

INITIAL EXPERIMENT			
volts (V)	degrees F (T)	% ontime	
80	74	35	
90	76	32	
100	79	30	
110	83	28	
120	88	25	
130	94	23	

The fitted equation is: % ontime = 52.3 - 0.25 V + 0.036 T
correlation coefficient = 0.998


 The experiment was repeated, with almost identical results.
 The last % ontime changed from 23 to 20. All other results are the same. . .

REPEATED EXPERIMENT			
volts (V)	degrees F (T)	% ontime	
80	74	35	
90	76	32	
100	79	30	
110	83	28	
120	88	25	
130	94	20	

The fitted equation is: % ontime = 77.9 - 0.08 V + 0.50 T
Correlation coefficient = 0.994

3. Relationships between independent variables (multi-collinearity) (cont'd)

The 2 sets of data are almost identical, but the fitted equations are quite different:

$$1. \% \text{ ontime} = 52.3 - 0.25 V + 0.036 T$$

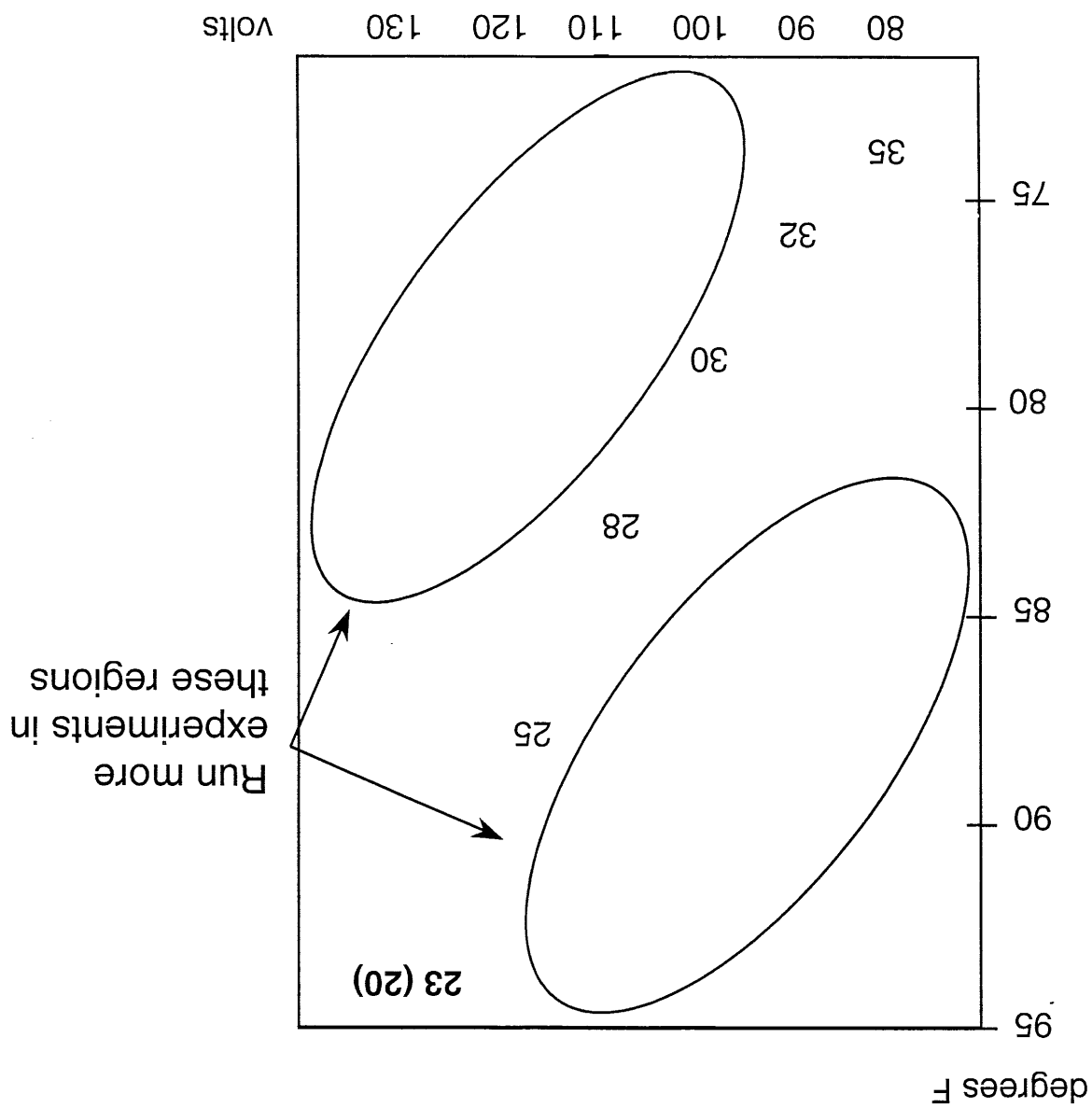
$$2. \% \text{ ontime} = 77.9 - 0.08 V + 0.50 T$$

The two independent variables, Volts and Temperature, are correlated. They change together, and it is not possible to determine if the change in the response is due to volts, temperature, or both.

Both equations will give nearly the same prediction of % ontime within the narrow band of voltage and temperature used in this experiment, but the predictions will be quite different for other combinations of voltage and temperature. (Look at the graph on the next page)

3. Relationships between independent variables (multi-collinearity) (cont'd)

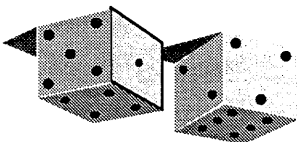
Plotted value is % *ontime*



Cannot separate effects caused by Voltage and Temperature changes. Need to take more data in circled regions...

4. Over-fitting; multiple hypothesis tests; too many independent variables

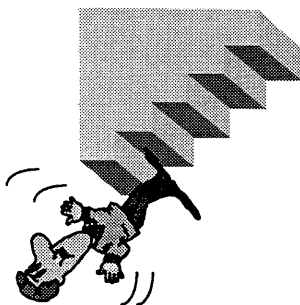
If a large number of independent variables are considered, then one would expect some of them to appear to be related to the dependent variable due to chance alone.



Number of independent variables
Probability that at least 1 is significant at the 95% confidence level

1	.05
2	.10
3	.14
4	.19
5	.23
10	.40
20	.64
30	.79
40	.87

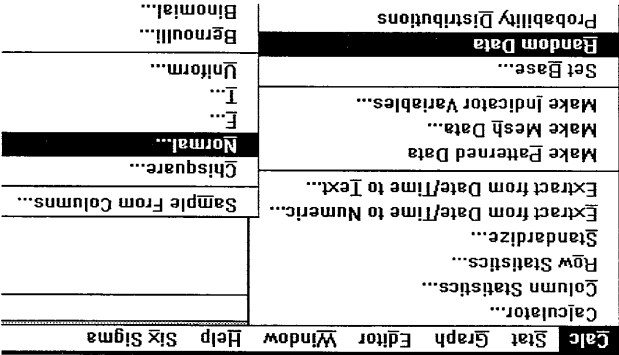
'Stepwise' regression is sometimes used to choose which of a large number of variables provides the best prediction. This can result in an over-fitted model (too many independent variables), and can give poor predictions of future values.



Let's try it! We'll generate random data, and see if there are significant correlations

In Minitab, create 200 rows and 30 columns of random normal data. Label C1 as "response", and use the remaining columns as the predictors ("Xs").

Calc>Random Data>Normal



Fill in the dialog box as shown:

Normal Distribution

Generate 200 rows of data

Store in column(s): C1 - C30

Mean: 0.0

Standard deviation: 1.0

Select

Help

OK

Cancel

Since we generated random data, your numbers will be different!

MINITAB - Untitled Worksheet - [Data]									
File Edit Manip Calc Stat Graph Editor Window Help Six Sigma									
	C1	C2	C3	C4	C5	C6	C7		
↑	Response	X1	X2	X3	X4	X5	X6		
1	-0.37820	-0.62119	1.32327	-0.57171	-0.44734	-0.81640	-1.26537	0.	
2	-1.69148	-1.24454	2.83157	0.35813	1.24798	0.22406	0.78629	-1.	
3	0.83943	1.72170	-0.37005	0.82515	1.30104	-0.89343	-0.47946	0.	
4	-1.06087	-1.02550	-0.76707	-0.35094	0.01969	-0.81201	-0.07859	0.	
5	-0.12768	-1.45869	-0.20969	-0.01224	-1.52448	-0.00622	1.03923	-0.	
6	0.22051	0.02722	-0.89498	-0.88438	1.34383	1.13363	-0.34500	0.	
7	-0.35086	-0.52063	1.51015	-0.16332	-0.24852	1.02796	-0.95034	1.	
8	0.85632	1.36410	0.68190	-1.18489	-1.44741	0.77965	0.18726	-0.	
9	-0.39705	1.13073	-0.56481	1.68087	-0.19869	-2.25412	-1.09236	-0.	
10	1.42892	-1.69785	-0.27899	0.72151	0.42917	-1.97142	1.00156	0.	
11	0.37141	1.19248	1.07837	-0.30517	-0.20298	-1.31436	-0.06186	1.	
12	0.18905	0.57197	0.31206	-0.21634	-0.21449	-0.87935	0.21828	1.	

Next, let's do regression analysis on this totally random data

Stat>Regression>Regression

Response: C1

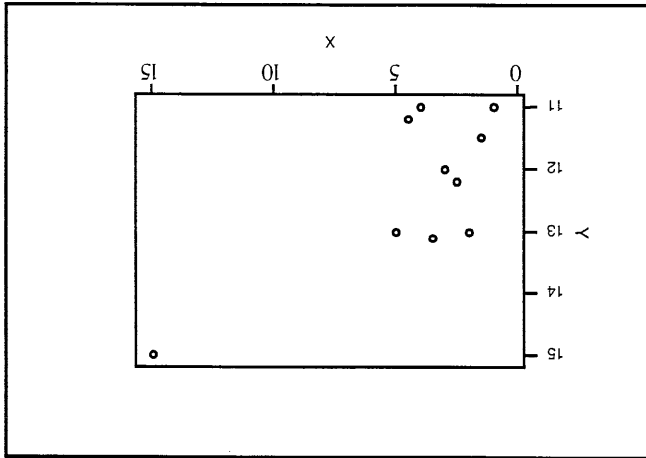
Predictors: C2 - C30

The screenshot shows the Minitab Regression dialog box. The 'Response' field is set to 'Response' and the 'Predictors' field is set to 'X1-X29'. A list of variables (C13 to C30) is shown in the 'Select' list. The dialog includes buttons for 'Help', 'Select', 'Options...', 'Storage...', and 'OK'.

Low p-values indicate factors that significantly influence Y. How many of your "predictors" have low p-values ($< .05$)?

5. Influence of a few extreme values

BOW	X	Y
1	1.0	11.0
2	1.5	11.5
3	2.0	13.0
4	2.5	12.2
5	3.0	12.0
6	3.5	13.1
7	4.0	11.0
8	4.5	11.2
9	5.0	13.0
10	15.0	15.0



All 10 observations:
Correlation of X and Y = 0.758

First 9 observations only:
Correlation of X and Y = 0.208

The apparent relationship is due almost entirely to the influence of the 10th observation.

Investigate the "Extreme Value". If it appears to be valid, then get more data at $X = 15$. DON'T AUTOMATICALLY THROW OUT THE "EXTREME" DATA POINT!!

6. Happenstance Data (Observational studies -- not designed experiments)

Happenstance data is sometimes called **PARC** data:

Post
Analysis
Regression and
Correlation

or

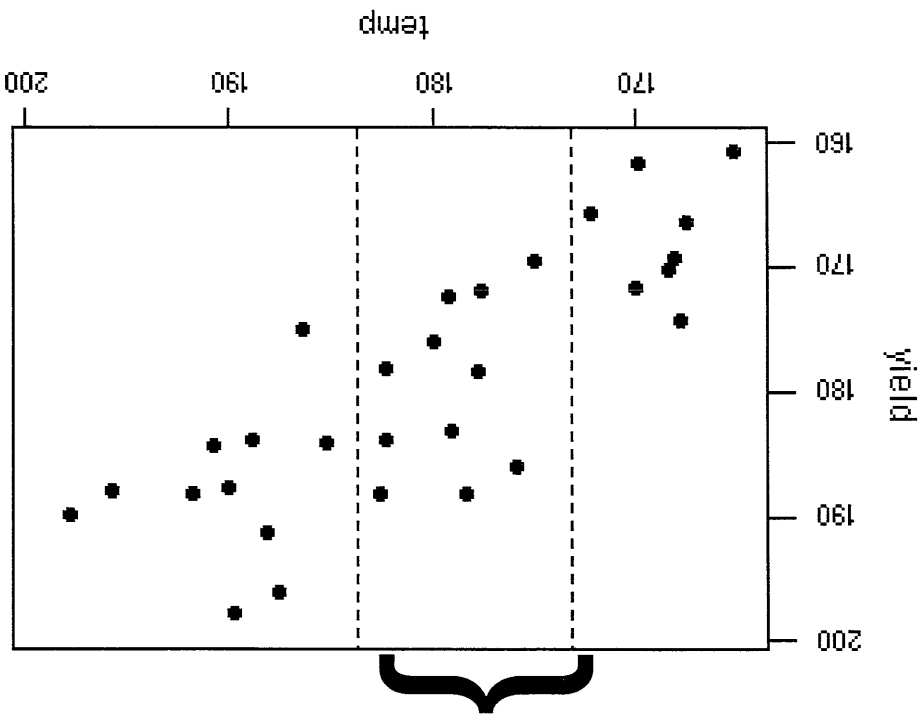
Plan
After
Research
Completed

Read it Backwards!

6. Happpenstance Data

A) Small range of the "X" variable

Observed range
of temperatures in
our experiment



A higher temperature could be used to increase yield.
However, we may not find this relationship since
temperature is tightly controlled in our sample.

6. Happpenstance Data

B) Lurking Variables

Another variable may systematically change during the study

- Temperature and humidity increase during the day
- Interest rates change, and may affect appliance sales.

Example:

A rotary compressor line was shut down early in the summer to add automation. The line was started again later in the summer, and the instrument measurements of noise increased.

The original assumption was that the process changes increased the noise.

It was later discovered that the noise measurement was not adequately controlled. The humidity increased during the summer, and this increased the measured noise.

The automation was not the problem.

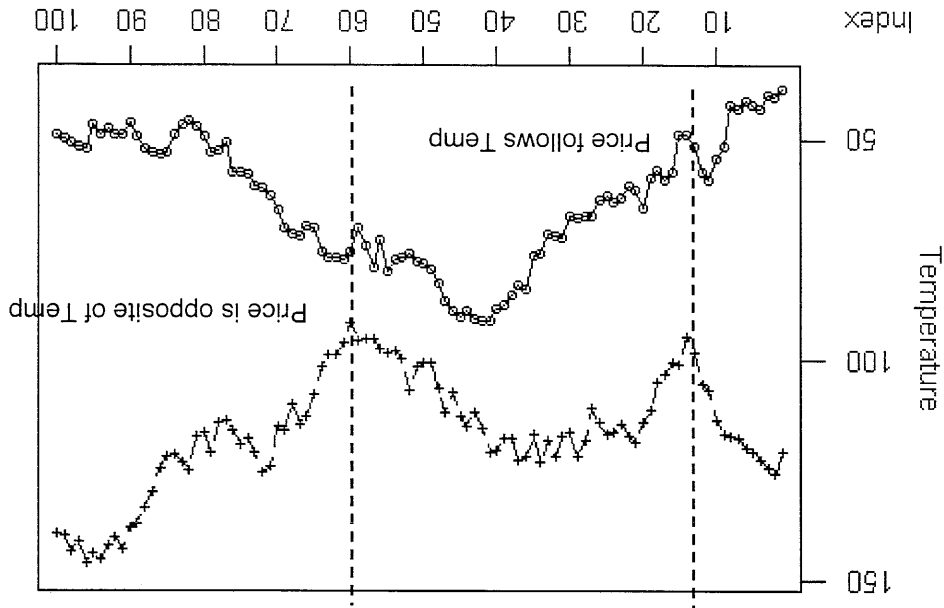
Humidity was a **lurking** variable.

6. Happpenstance Data

C) Variables that are "serially correlated" over time

Data that are collected over time are often "serially correlated". That is, data over a short period of time are similar, and data far apart in time are not as similar.

Louisville Temperatures and Stock price



The two data sets in this graph (local temperatures and Stock prices) were generated independently, and are not related over a long period of time.

However, there are short periods of time where they increase together or decrease together.

7. Always graph data before regression analysis

The following data can be found in: Edward R. Tufte. The Visual Display of Quantitative Information. Graphics Press.
Originally published in: F. J. Anscombe. "Graphs in Statistical Analysis." American Statistician. 27 (February, 1973), 17-21.

X1	5	7	12	4	6	14	11	9	13	8	10	8.04
Y1	5.68	4.82	10.84	4.26	7.24	9.96	8.33	8.81	7.58	6.95	8.04	
X2	5	7	12	4	6	14	11	9	13	8	10	9.14
Y2	4.74	7.26	9.13	3.10	6.13	8.10	9.26	8.77	8.74	8.14	9.14	
X3	5	7	12	4	6	14	11	9	13	8	10	7.46
Y3	5.73	6.42	8.15	5.39	6.08	8.84	7.81	7.11	12.74	6.77	7.46	
X4	8	8	8	19	8	8	8	8	8	8	8	6.58
Y4	6.89	7.91	5.56	12.50	5.25	7.04	8.47	8.84	7.71	5.76	6.58	

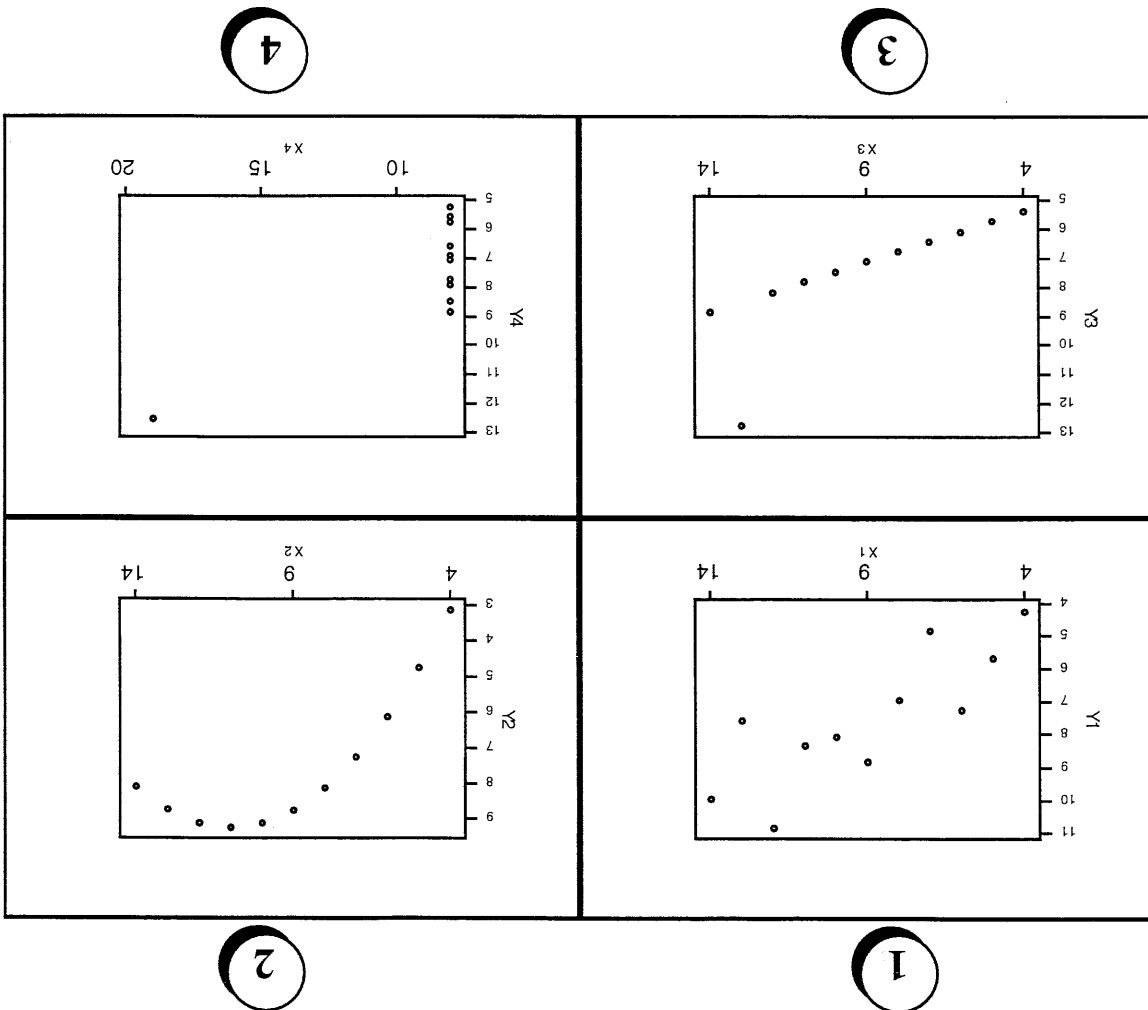
X1, X2, X3, and X4 have identical averages and standard deviations.
Y1, Y2, Y3, and Y4 have identical averages and standard deviations.
Regression for the 4 data sets give:

- Identical fitted equations
- Identical correlation coefficients
- Identical estimates of the standard deviation of the residuals
- Identical tests of significance for the slopes

The data are graphed on the following page.

The graphs show that the relationships are dramatically different, although the regression outputs are the same.

Always graph the data first - here's why:



Data Set #1: Line is valid.

Data Set #2: Fit a quadratic equation

Data Set #3: Investigate the one high point

(possibly bad data)

Data Set #4: Get more data for large values of X

The pitfalls of regression can be avoided through:

1. Use of theoretical models, and knowledge from other sources.

2. Proper data collection
 - Use experimental designs that are orthogonal
 - Use a wide range of the independent variables
 - Randomize the order of data collection over time

3. Use regression diagnostics
 - Scatter plots
 - Plots of residuals (observed - predicted)
 - ◆ versus predicted
 - ◆ versus independent variables
 - ◆ versus order

- Minitab will flag observations with large residuals and large influence -- Look at them!
- Significance tests for the estimated coefficients
- Standard deviation of the residuals (How good is the fit?)

4. Center independent variables when fitting squared terms and interactions (Response Surface Designs, for example).
Use:

Squared Terms: $(X_i - \bar{X})^2$

Interactions: $(X_{11} - \bar{X}_1) * (X_{21} - \bar{X}_2)$

Key Concepts: Pitfalls of Regression

1. Plot your data!!!

2. Check your assumptions

3. Get help when you need it