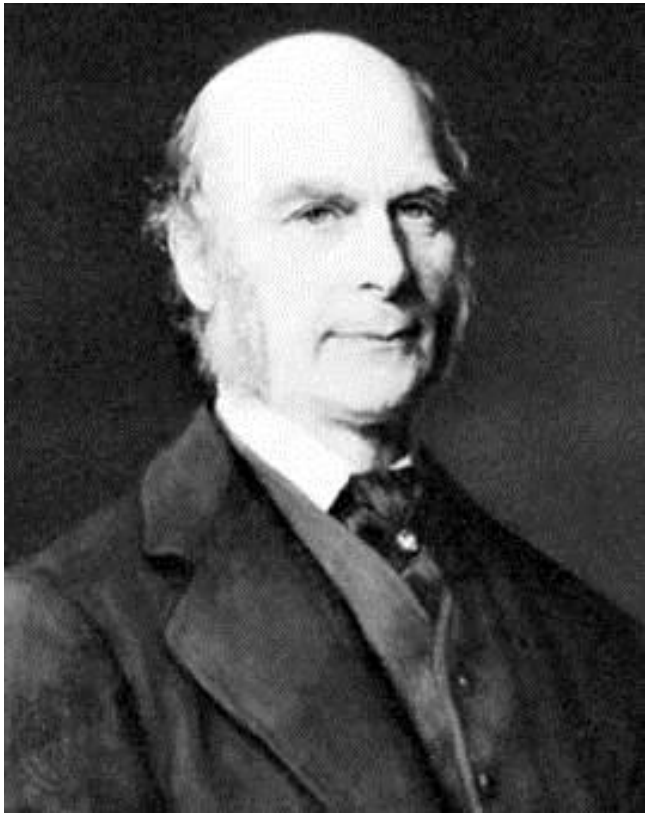
Statistical Science 1024

Chapter 5

Regression

Origin of the Term “Regression”

- Francis Galton, 1886, ‘Regression towards mediocrity in hereditary stature.’ *Journal of the Anthropological Institute*, **15**: 246 – 263
- See JSTOR under UWO library databases

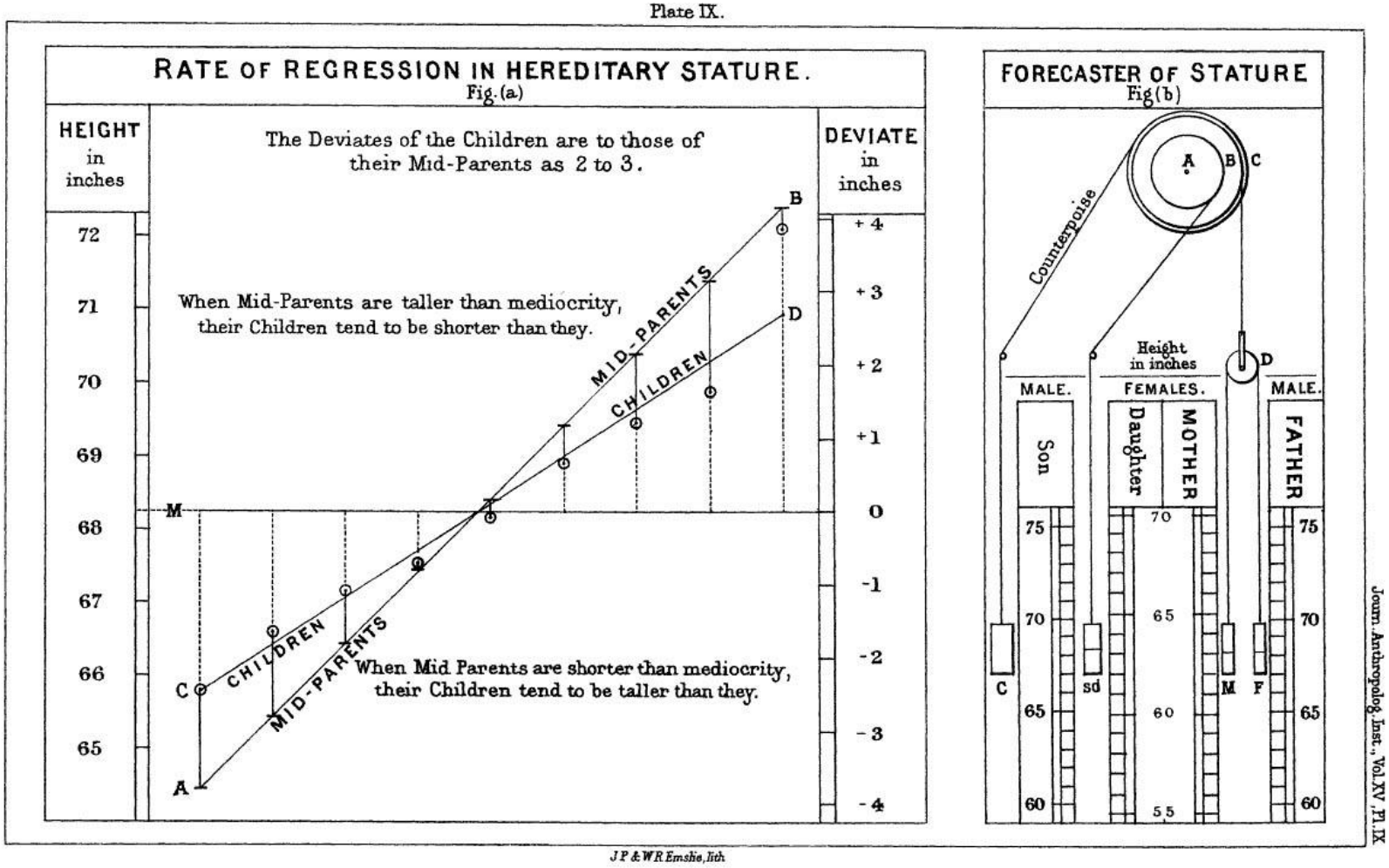


Data on Heights of Children and Parents

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

‘Regression Line’



REGRESSION LINE

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

REVIEW OF STRAIGHT LINES

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = a + bx$$

In this equation, b is the **slope,** the amount by which y changes when x increases by one unit. The number a is the **intercept,** the value of y when $x = 0$.

Example
Radioactive Waste and the Columbia River

- the Hanford Atomic Energy Plant in Washington State in the United States had been producing plutonium since World War II
- some of the waste from the plant was stored in open pits and the radioactive waste had leaked into the Columbia River
- eight Oregon counties and the city of Portland have been exposed to radioactive contamination

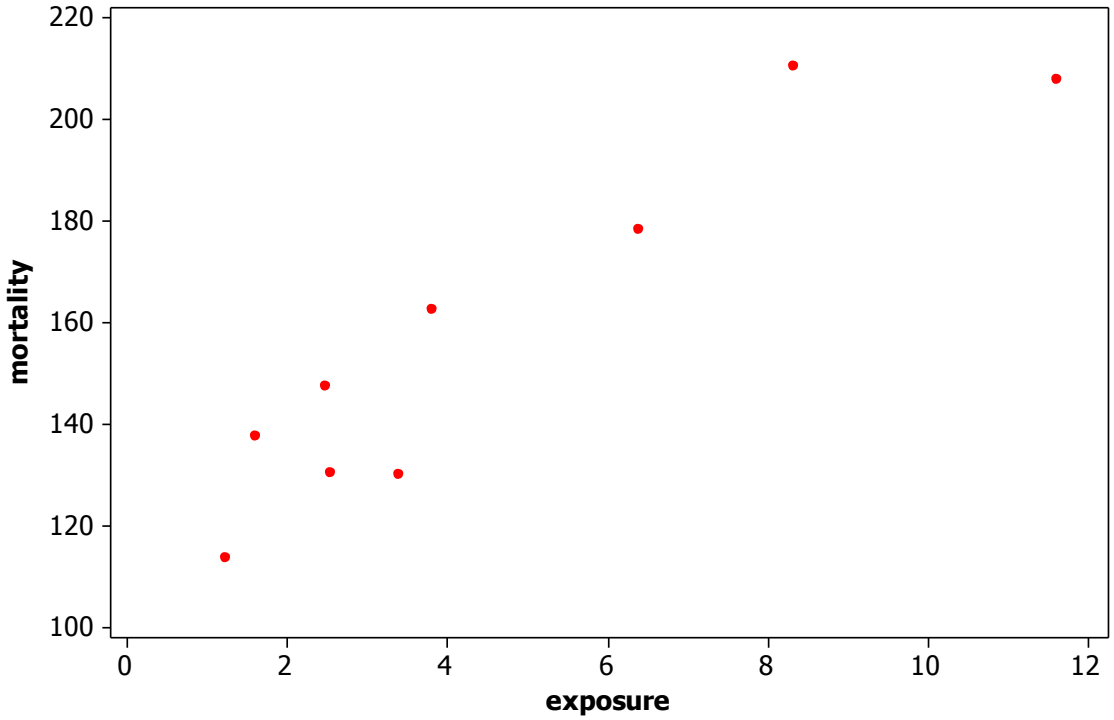
Data

County or city	Exposure Index	Cancer Mortality
Umatilla	2.49	147.1
Morrow	2.57	130.1
Gilliam	3.41	129.9
Sherman	1.25	113.5
Wasco	1.62	137.5
Hood River	3.83	162.3
Portland	11.64	207.5
Columbia	6.41	177.9
Clatsop	8.34	210.3

- original data collected in 1965 based on experience of the previous five years
- problem still exists today

Scatterplot

- the exposure index includes factors such as distance of the county from Hanford and the average distance from water frontage
- the cancer mortality rate is deaths per 100,000 residents

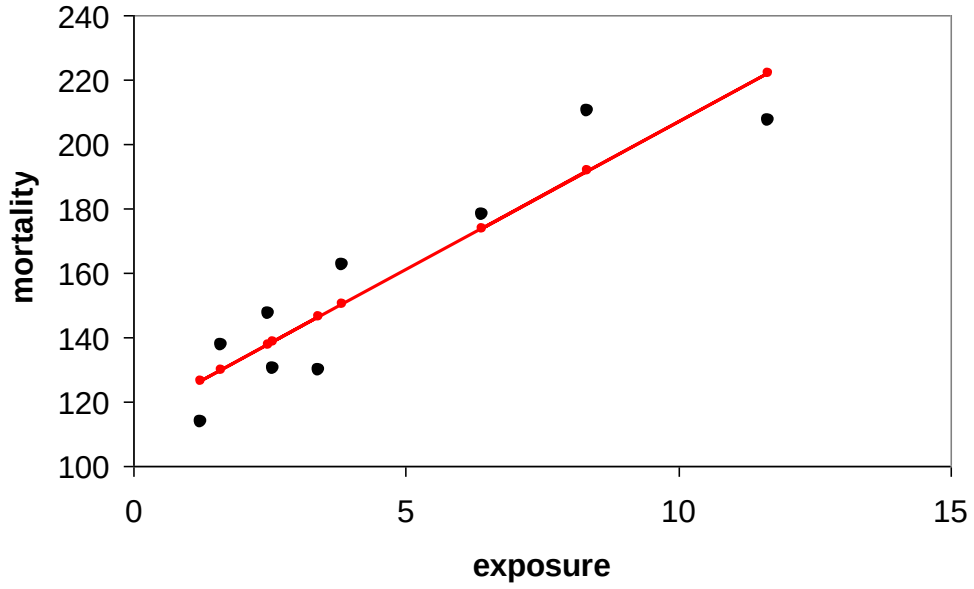


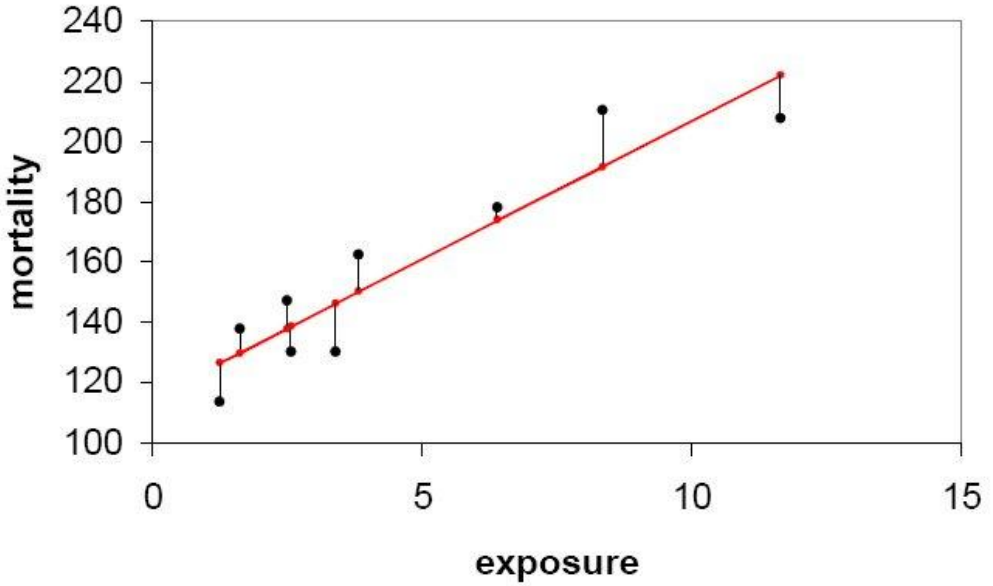
LEAST-SQUARES REGRESSION LINE

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Least-squares Regression Line

"the least-squares regression line of *y* on *x* is that line that makes the sum of squares of vertical distances of the data points from the line as small as possible"





EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\overline{x}$ and $\overline{y}$ and the standard deviations s_x and s_y of the two variables, and their correlation r. The least-squares regression line is the line

$$\hat{y} = a + bx$$

with **slope**

$$b = r\frac{s_y}{s_x}$$

and **intercept**

$$a = \overline{y} - b\overline{x}$$

Calculation of the Estimates

$$\bar{x} = \frac{2.49 + 2.57 + \cdots 8.34}{9} = 4.62$$

$$\bar{y} = \frac{147.1 + 130.1 + \cdots + 210.3}{9} = 157.3$$

$$s_x = \sqrt{\frac{(2.49 - 4.62)^2 + \cdots + (8.34 - 4.62)^2}{8}} = 3.49$$

$$s_y = \sqrt{\frac{(147.1 - 157.3)^2 + \cdots + (210.3 - 157.3)^2}{8}} = 34.8$$

$$r = \frac{(2.49 - 4.62)(147.1 - 157.3) + \cdots + (8.34 - 4.62)(210.3 - 157.3)}{8(3.49)(34.8)} = 0.926$$

$$b = 0.926\frac{34.8}{3.49} = 9.23$$

$$a = 157.3 - (9.23)(4.62) = 114.7$$

Always plot your data!

Here are four data sets. The correlations all give $r \approx 0.816$, and the regression lines are all approximately $\hat{y} = 3 + 0.5x$. For all four sets, we would predict $\hat{y} = 8$ when $x = 10$.

Data Set A

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

Data Set B

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C

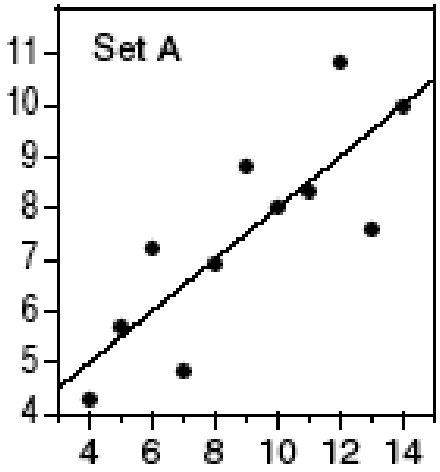
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D

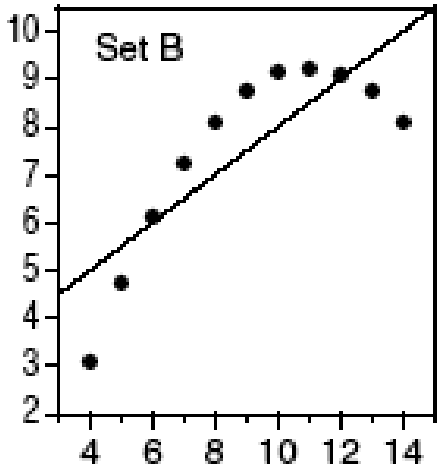
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Always plot your data!

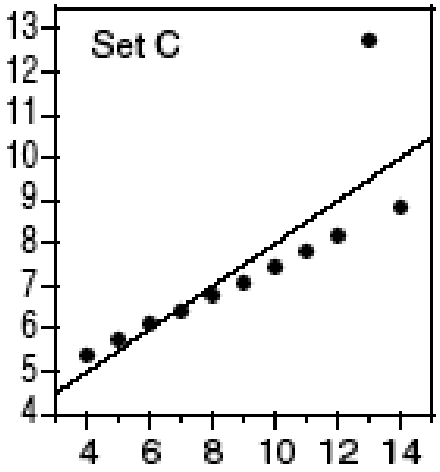
However, making the scatterplots shows us that the correlation/ regression analysis is not appropriate for all data sets.



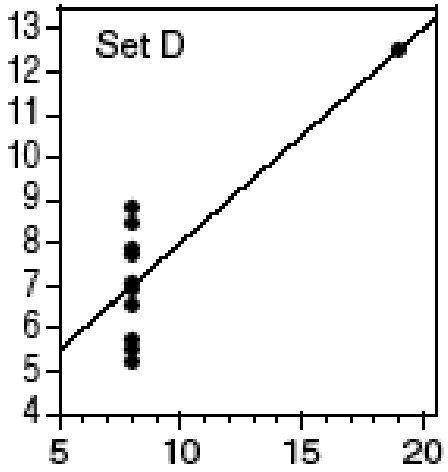
Moderate linear association; regression OK.



Obvious nonlinear relationship; linear regression inappropriate.



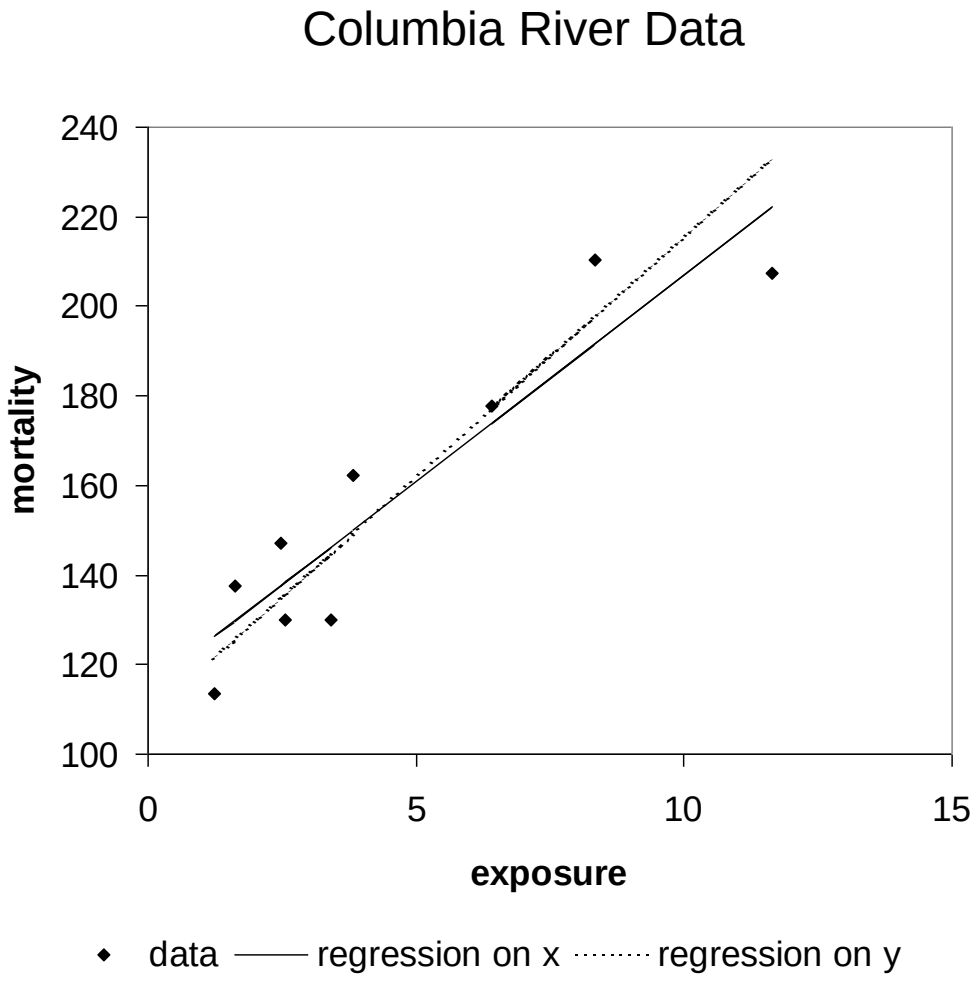
One point deviates from the (highly linear) pattern of the other points; it requires examination before a regression can be done.



Just one very influential point and a series of other points all with the same *x* value; a redesign is due here…

Regression Fact 1

- the distinction between explanatory and response variables is essential
- regression of *y* on *x* is different from regression of *x* on *y*

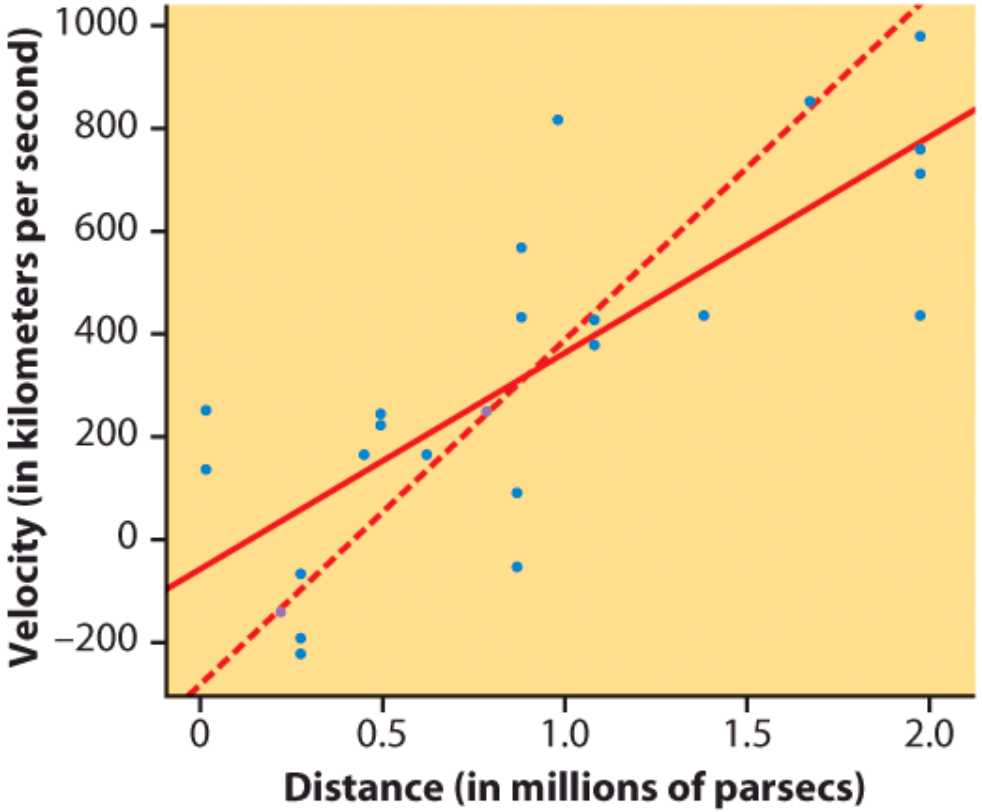


Another Example

Regression examines the distance of all points from the line <u>in the y direction only</u>.

Data from the Hubble telescope about galaxies moving away from Earth:

These two lines are the two regression lines calculated either correctly (x = distance, y = velocity, solid line) or incorrectly (x = velocity, y = distance, dotted line).



Regression Fact 2

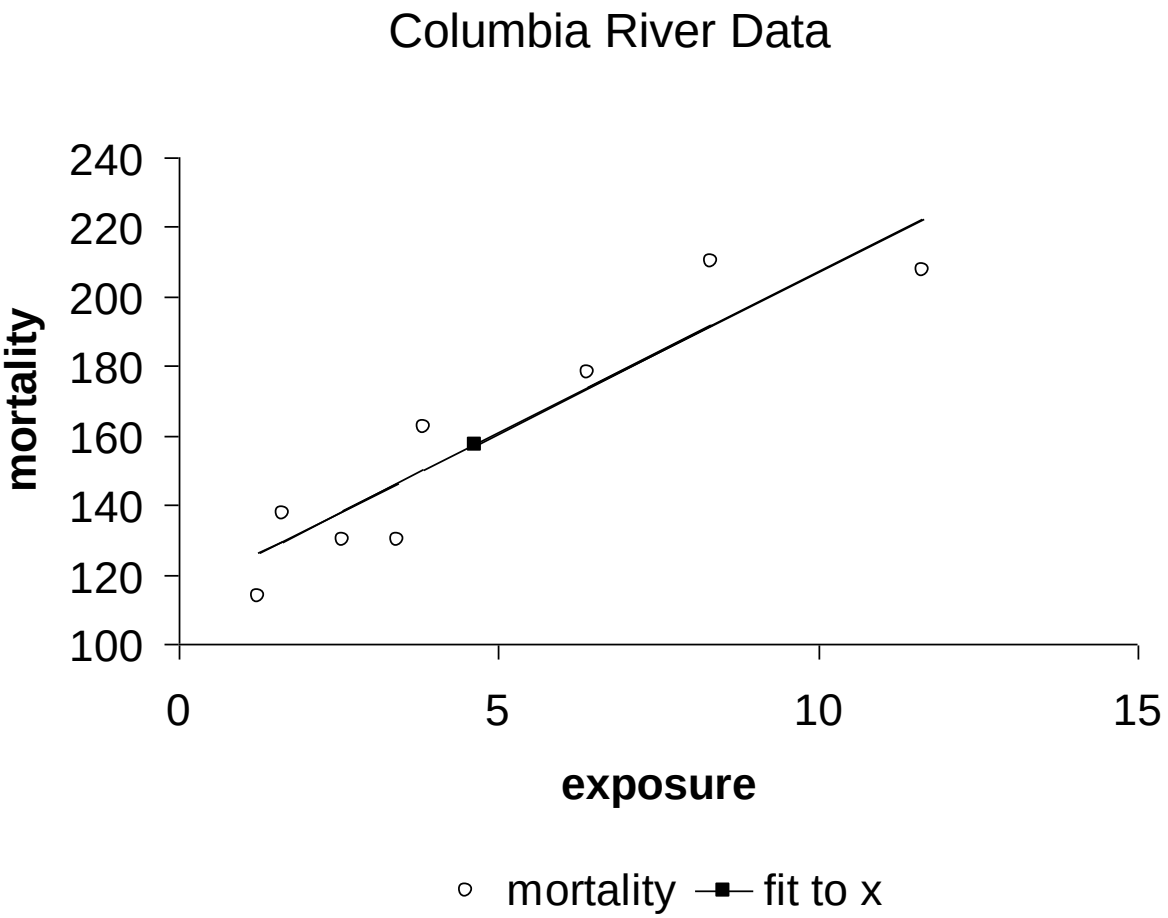
$$\hat{y} = a + bx \qquad b = r\frac{s_y}{s_x}$$

- a change of one standard deviation in *x* corresponds to a change of *r* standard deviations in *y*

Regression Fact 3

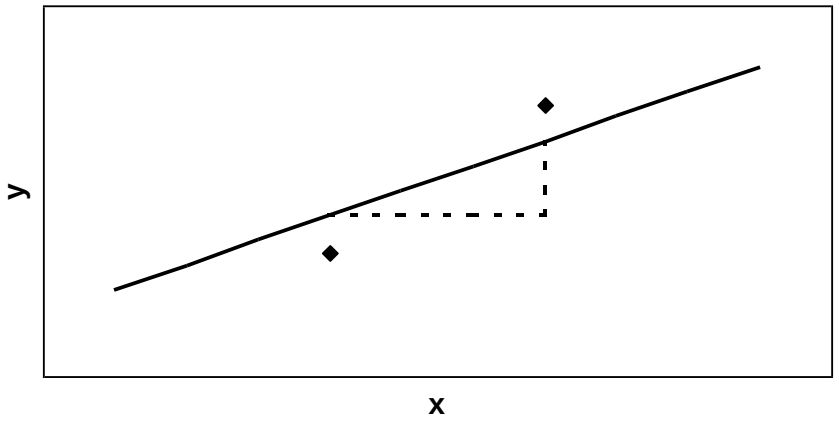
- the regression line passes through the point: (4.62,157.3)

$$(\bar{x}, \bar{y})$$



Regression Fact 4

- the square of the correlation, i.e. r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of *y* on *x*
- two types of variation: (1) due to the line and (2) about the line.

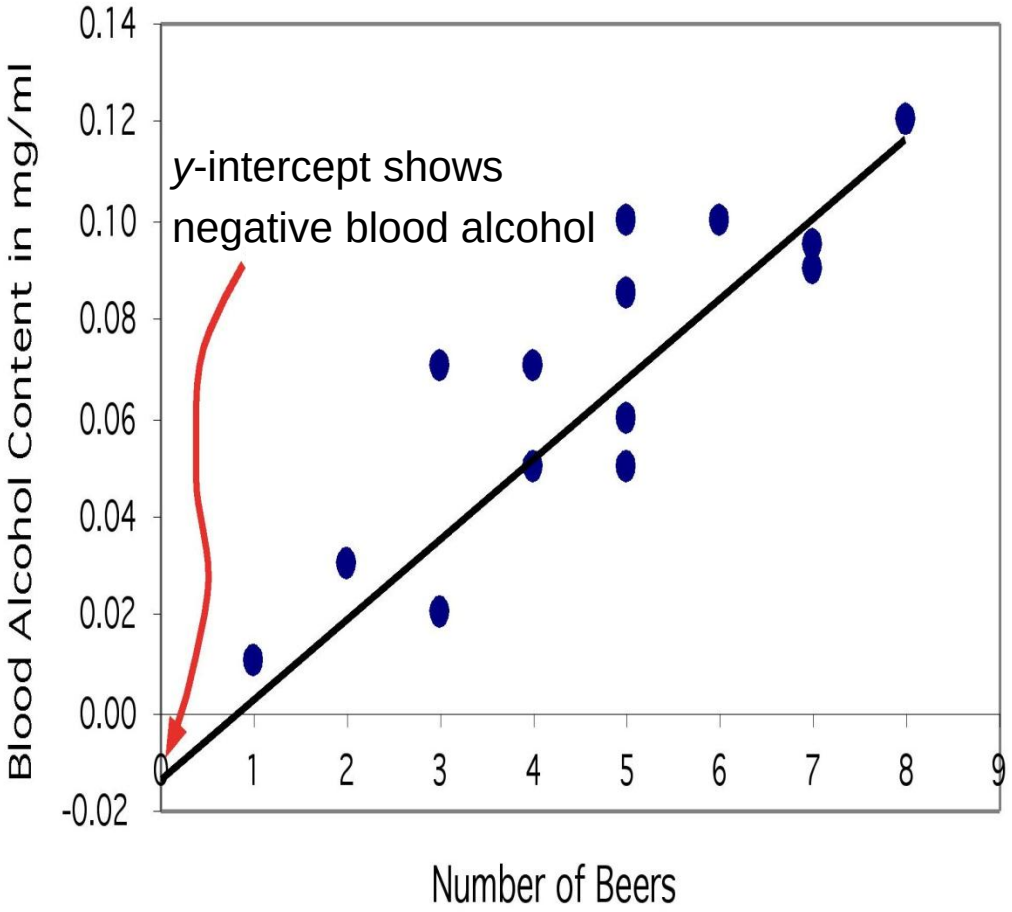


The y-intercept

Sometimes the *y*-intercept is not biologically possible. Here we have negative blood alcohol content, which makes no sense…

But the negative value is appropriate for the equation of the regression line.

Should use regression through the origin – set $a = 0$ and estimate b.



RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is, a residual is the prediction error that remains after we have chosen the regression line:

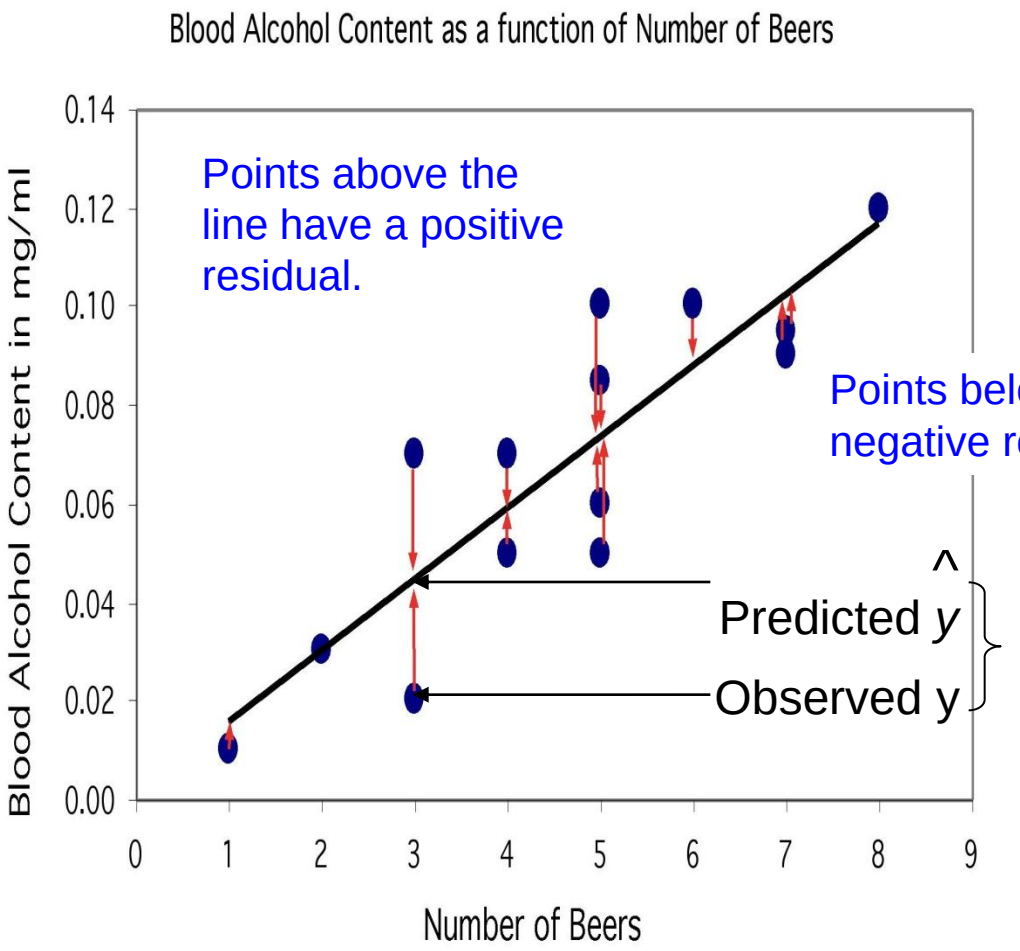
$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y}\end{aligned}$$

Residuals

The distances from each point to the least-squares regression line give us potentially useful information about the contribution of individual data points to the overall pattern of scatter.

These distances are called **"residuals."**

The sum of these residuals is always 0.



$$\text{dist. } (y - \hat{y}) = \text{residual}$$

Properties of Least-Squares Residuals

- residuals may be positive or negative
  - a positive value indicates the observed value lies above the regression line
  - a negative value indicates that it falls below the line
- the sum of residuals in least-squares regression is 0

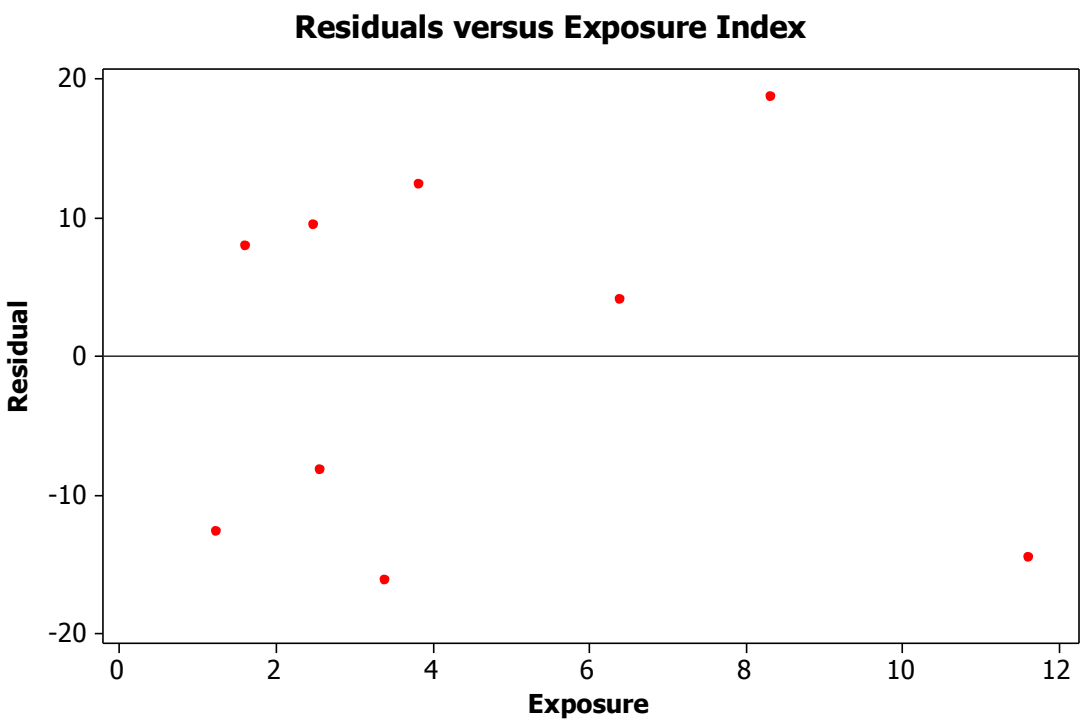
RESIDUAL PLOTS

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.

Residual Plots

- scatterplot of the regression residuals against the explanatory variable (*x*)
- used to help assess the fit of the regression line
- residual plot of the Columbia River contamination data

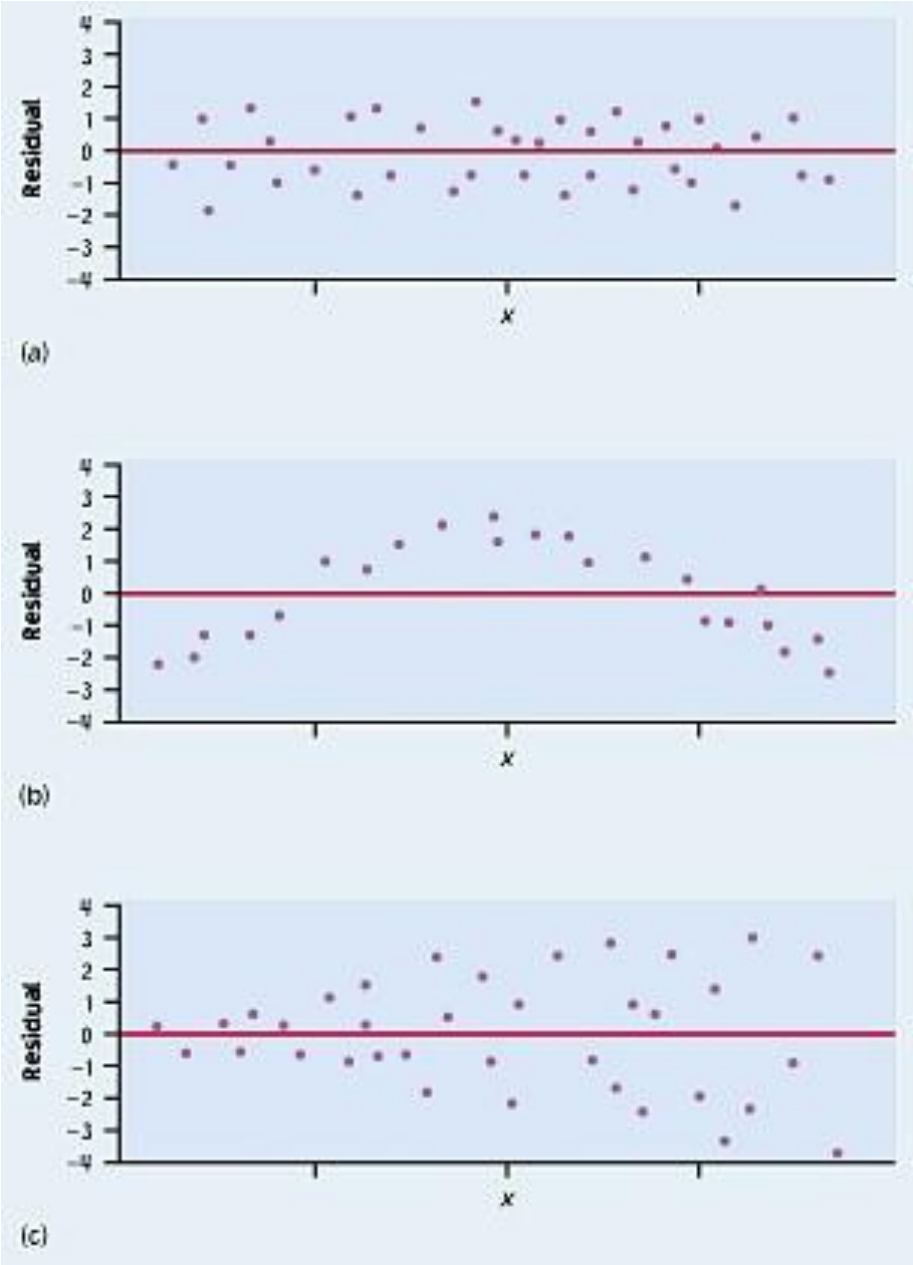
Columbia River Data



Examining Residual Plots

Apply the general principles of examining scatterplots by looking for:

- the overall pattern in the plot (ideally you would like to see no pattern in the residuals)
- the form and direction of any pattern if it exists
- deviations from the pattern (outliers and influential observations)

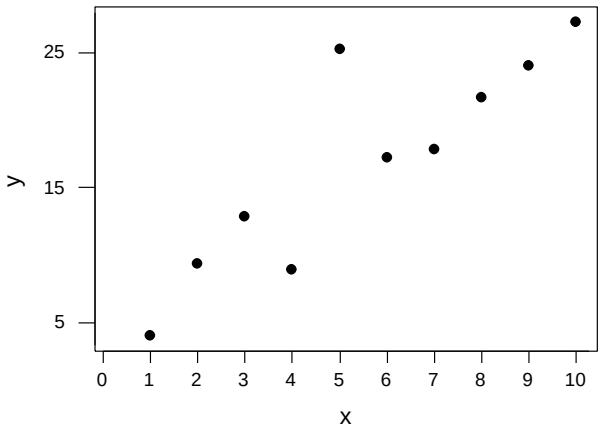


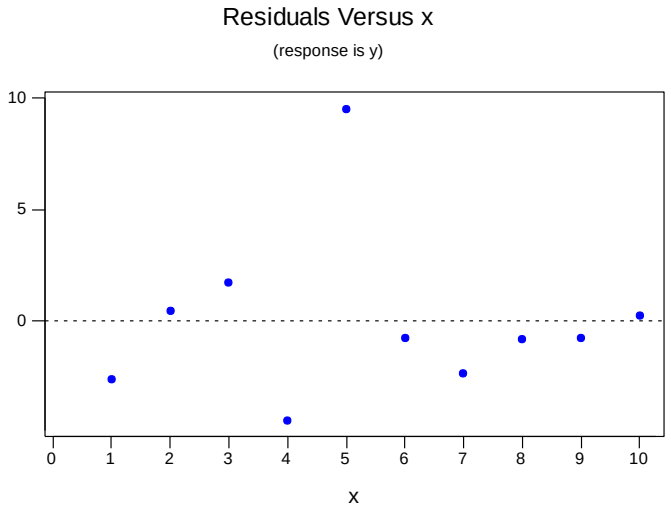
Residuals are randomly scattered—good!

A curved pattern—means the relationship you are looking at is not linear.

A change in variability across plot is a warning sign. You need to find out why it is and remember that predictions made in areas of larger variability will not be as good.

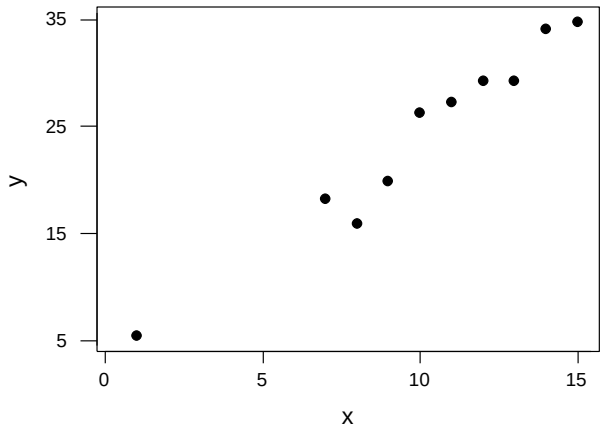
Points with Large Residuals

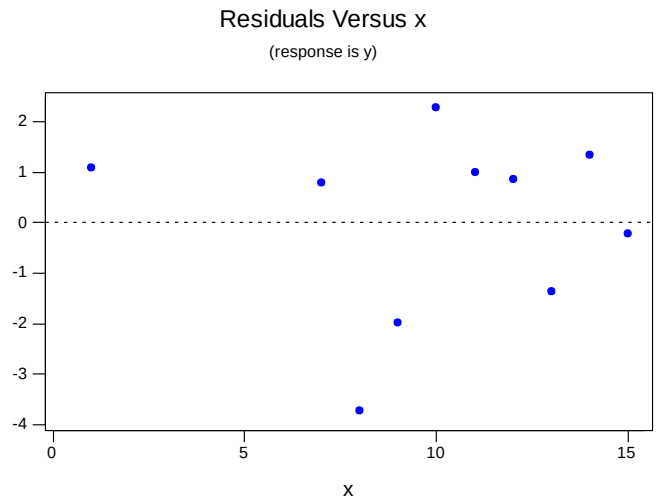




- points are outliers in the vertical direction because they lie far from the line that describes the overall pattern

Extreme Points in the *X* Direction





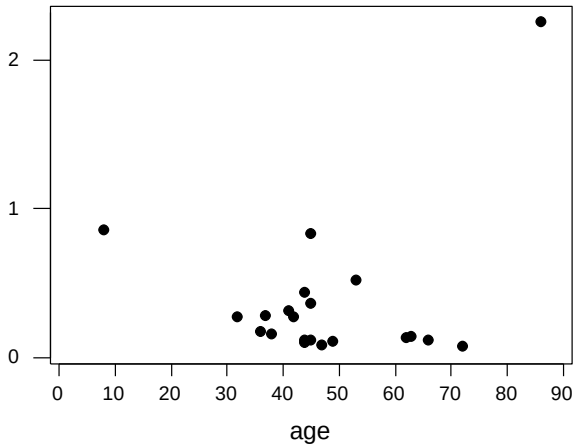
- the residual associated with the smallest *x* here is not a large residual but it could have a considerable influence on the fit of the line

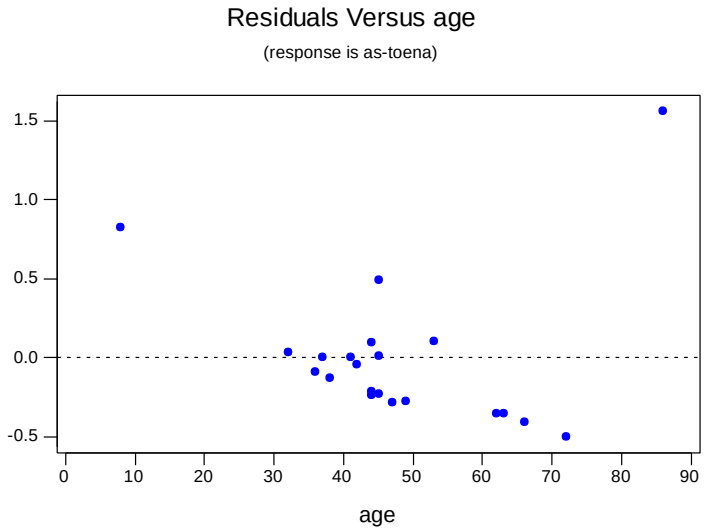
INFLUENTIAL OBSERVATIONS

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation.

Points that are outliers in either the x or y direction of a scatterplot are often influential for the correlation. Points that are outliers in the x direction are often influential for the least-squares regression line.

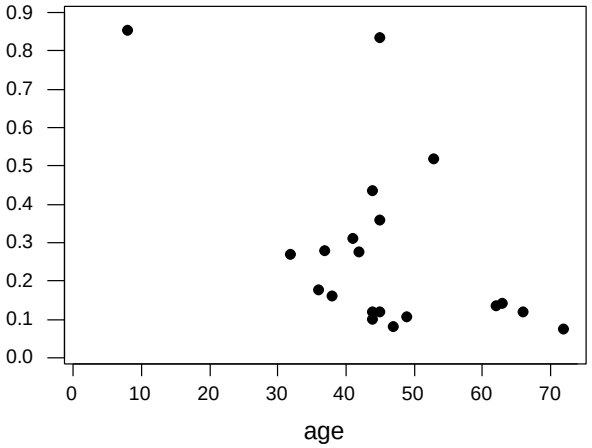
Outlier Example: Arsenic in Well Water and in the Body

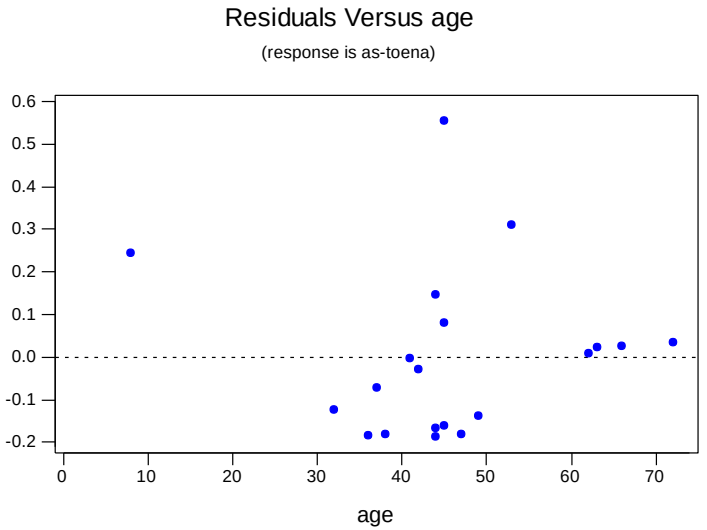




$$\hat{y} = -0.038 + 0.00850x, r^2 = 7.9\%$$

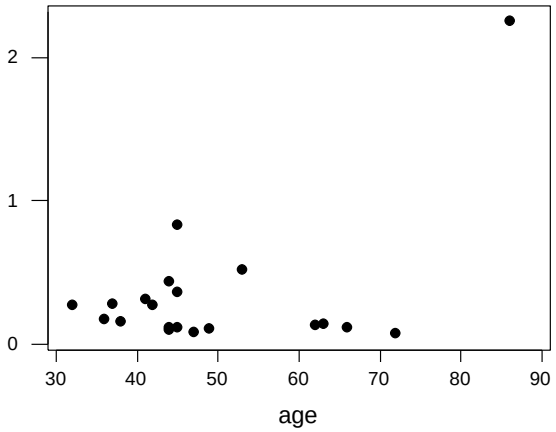
Older Age Removed

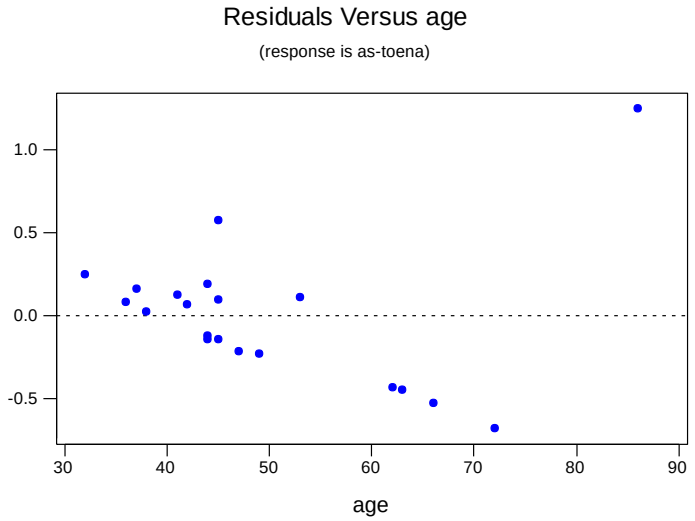




$$\hat{y} = 0.678 - 0.00888x,\ r^2 = 28.3\%$$

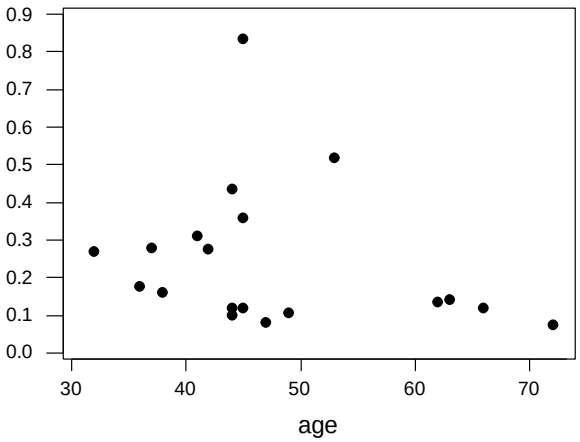
Younger Age Removed

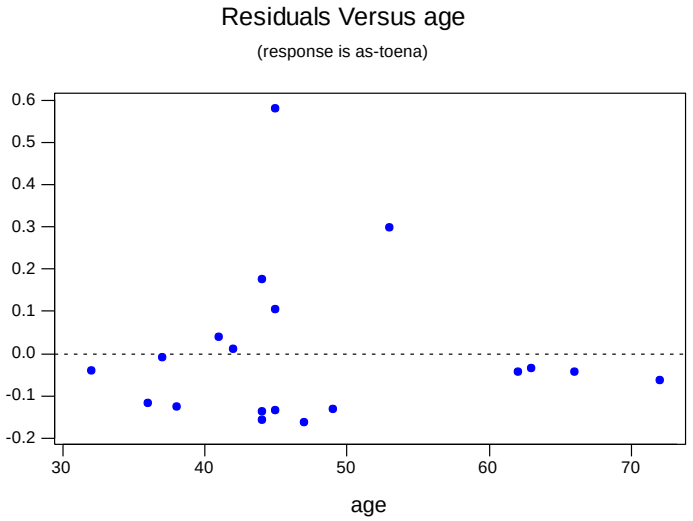




$$\hat{y} = -0.559 + 0.0182x,\ r^2 = 25.9\%$$

Both Removed

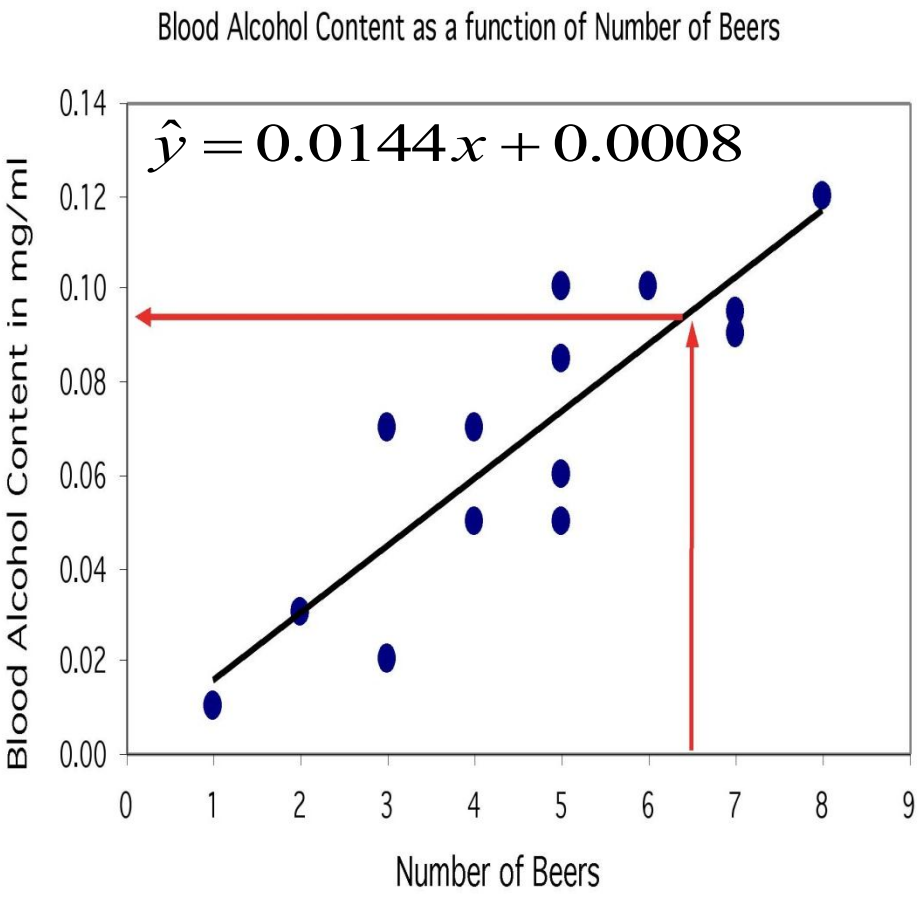




$$\hat{y} = 0.448 - 0.00432x, r^2 = 6.1\%$$

Making predictions: Interpolation

The equation of the least-squares regression allows you to predict *y* for any *x* **within the range studied.** This is called **interpolating**.



Nobody in the study drank 6.5 beers, but by finding the value of $\hat{y}$ from the regression line for *x* = 6.5, we would expect a blood alcohol content of 0.094 mg/ml.

$$\hat{y} = 0.0144(6.5) + 0.0008$$

$$\hat{y} = 0.936 + 0.0008 = 0.0944\ mg/ml$$

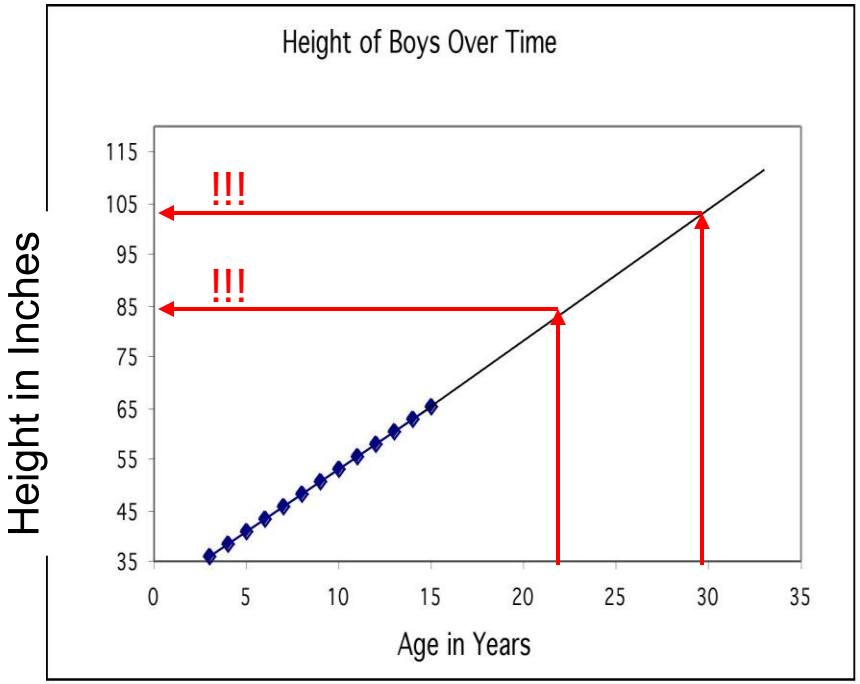
EXTRAPOLATION

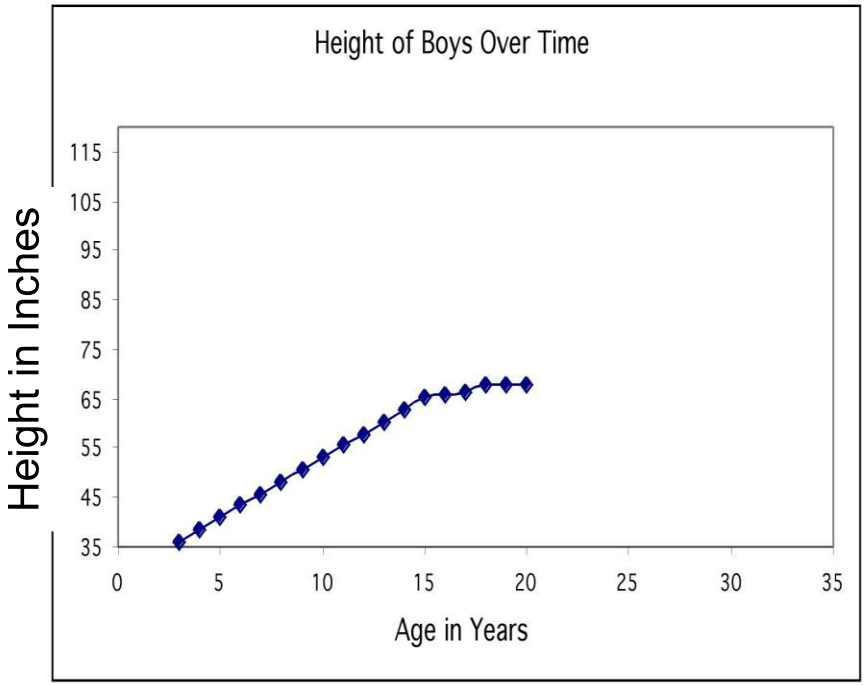
Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

More on Extrapolation

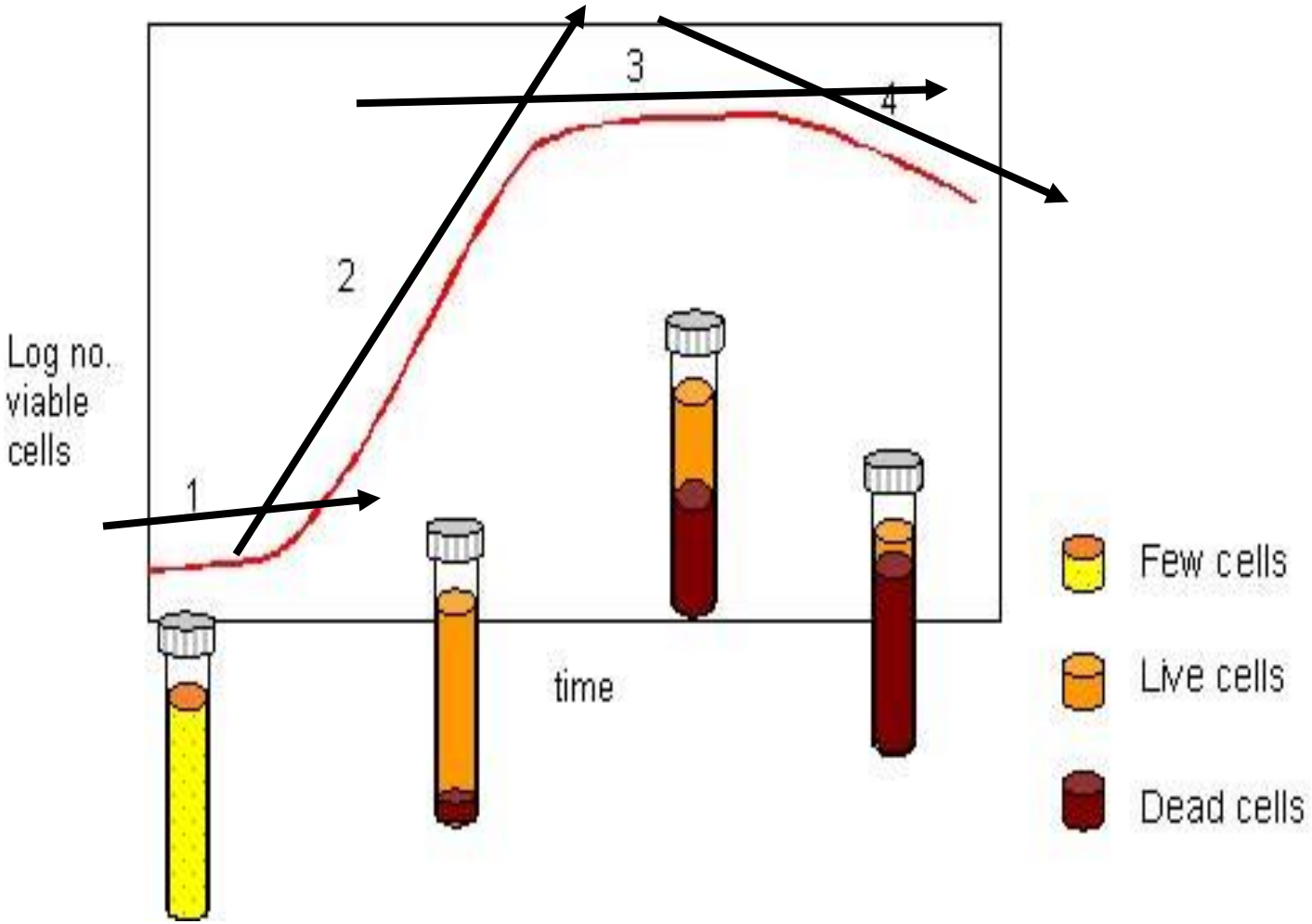
Extrapolation is the use of a regression line for predictions outside the range of *x* values used to obtain the line.

This can be a very stupid thing to do, as seen here.





Bacterial growth rate changes over time in closed cultures:



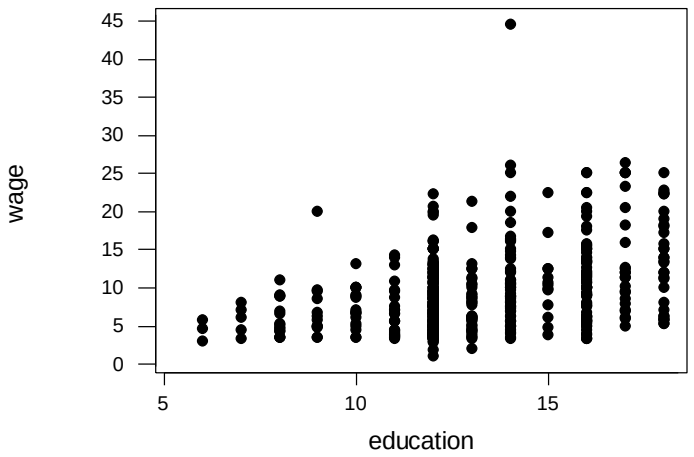
If you only observed bacterial growth in test tubes during a small subset of the time shown here, you could get almost any regression line imaginable.

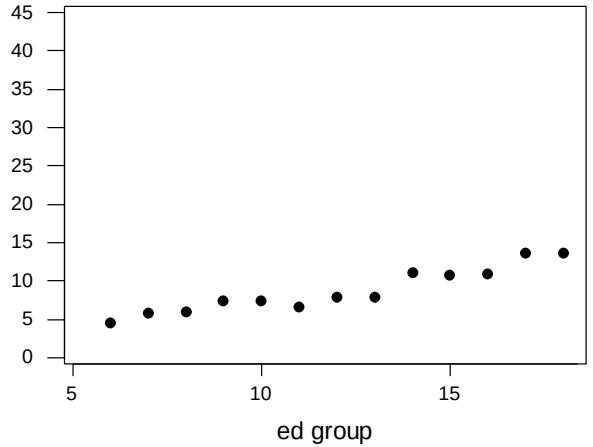
Extrapolation = big mistake

Using Averaged Data

- correlations based on averages are usually too high when applied to individuals
- example: Current Population Survey (1985) from the U.S.
  - responses were obtained for years of education and wages in dollars per hour on 530 individuals

1985 Current Population Survey





- raw data (left graph) $r = 0.394$
- averaged by years of education (right graph) $r = 0.952$

LURKING VARIABLE

A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

Lurking variables

A **lurking variable** is a variable not included in the study design that does have an effect on the variables studied.

Lurking variables can <u>falsely suggest</u> a relationship.

What is the lurking variable in these examples?

How could you answer if you didn't know anything about the topic?

• Strong positive association between the number firefighters at a fire site and the amount of damage a fire does

– Negative association between moderate amounts of wine drinking and death rates from heart disease in developed nations

Vocabulary: lurking versus confounding

LURKING VARIABLE

- A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

CONFOUNDING

- Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

But you often see them used interchangeably…

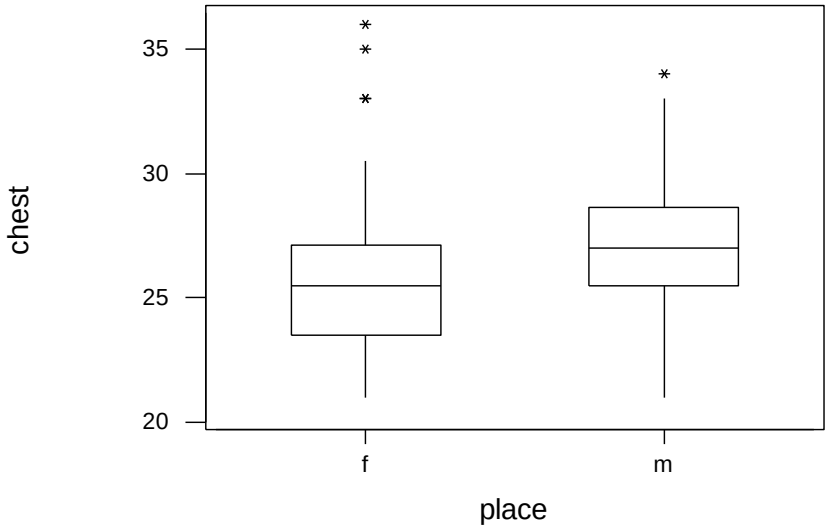
Back to Lurking Variables

- a **lurking variable** is a variable that has an important effect on the relationship among the variables in a study but is not included among the variables studied
- example: 1841-42 Children's Employment Commission of Great Britain – comparison of young girls working on farms and in mines

Quote from the Report

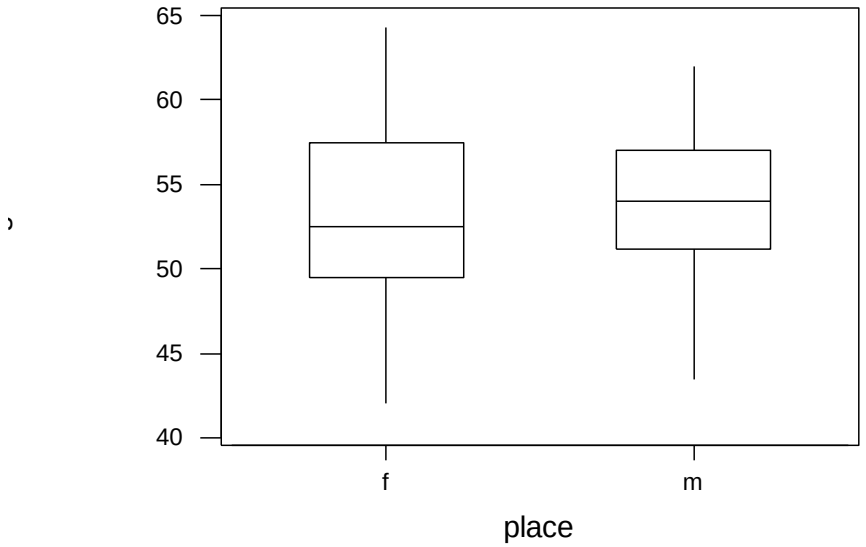
- “In stature there is an appreciable difference in colliers’ children [compared to farm children], manifest at all ages after they have been three years constantly in the pits: there is little malformation, but as Mr. Eliss, a surgeon constantly attending them, admits they are somewhat stunted in growth and expanded in width.”

Chest Measurements



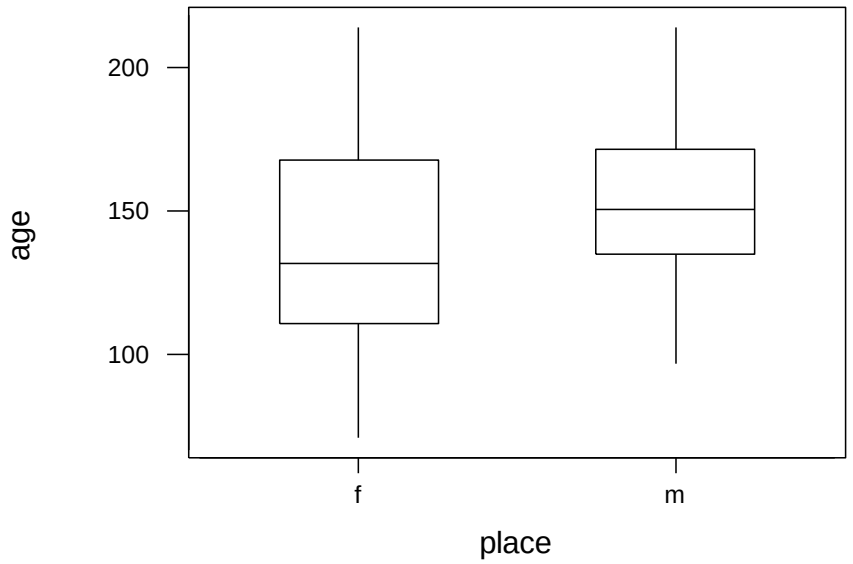
- Report: “colliers’ children … are somewhat … expanded in width”
- the boxplots agree with this assessment

Heights



- Report: “colliers’ children … are somewhat stunted in growth”
- the boxplots show the median height of collier’s children larger that farm children

Lurking (or Confounding) Variable: Age



- the boxplots show that colliers' children are older than farm children
- tend to be taller as a group because of age

Height Comparison by Age

Ages: miners	Mean height	Age: farmers	Mean height
		5	42
		6	46.75
		7	47.75
8	47.13	8	47.88
9	48.25	9	50.54
10	50.15	10	51.72
11	53.25	11	53.75
12	53.6	12	53
13	54.96	13	56.75
14	56.33	14	58.87
15	59.75	15	60.75
16	60	16	61.63
17	61	17	61

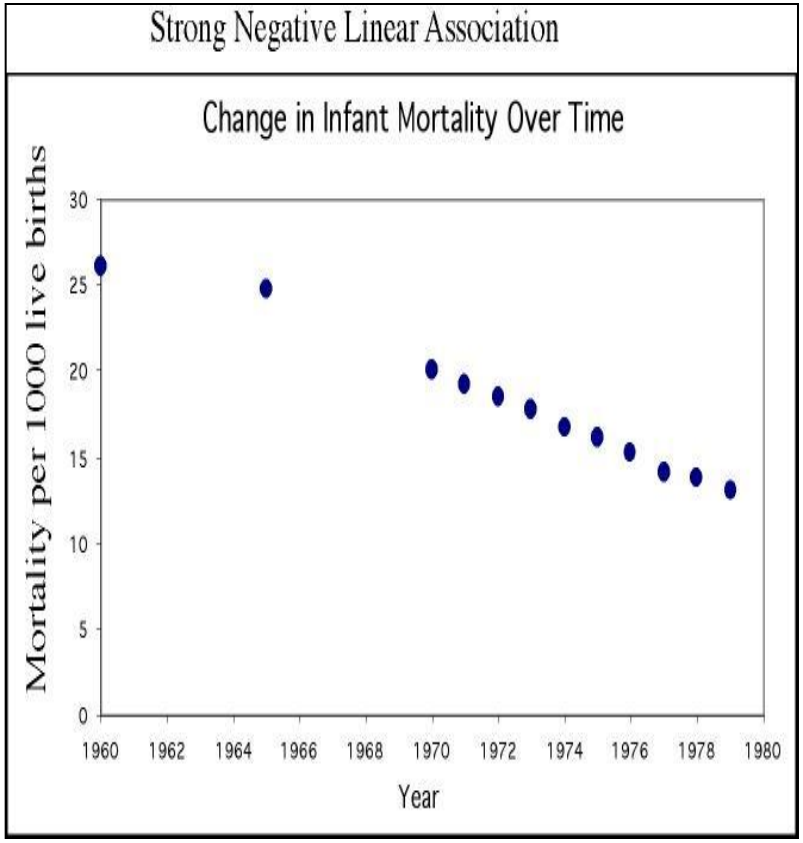
ASSOCIATION DOES NOT IMPLY CAUSATION

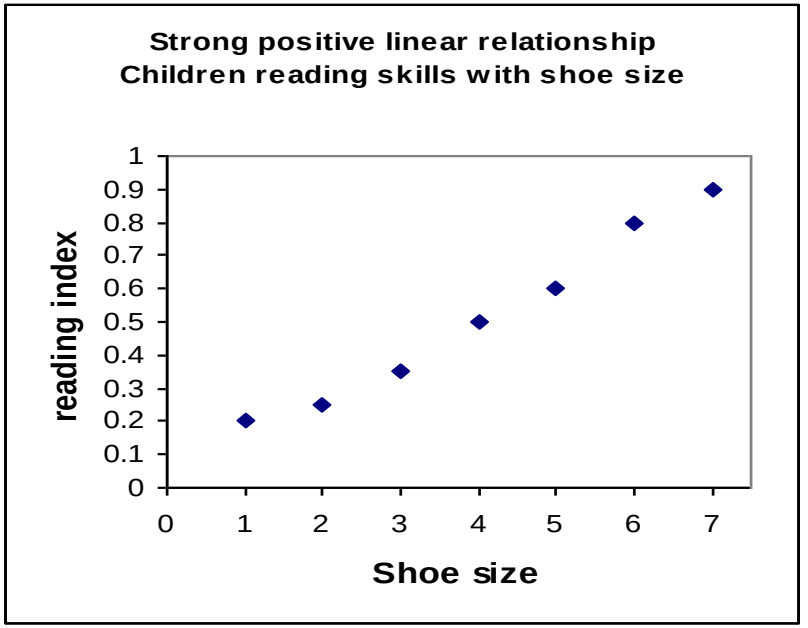
An association between an explanatory variable x and a response variable y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.

Association and causation

Association, however strong, does NOT imply causation.

Only careful experimentation can show causation.





Not all examples are so obvious…

Determination of Cause

- carry out a designed experiment by manipulating the explanatory variable *x* to see what happens to the response variable *y*

Association and causation

It appears that lung cancer is associated with smoking.

How do we know that both of these variables are not being affected by an unobserved third (lurking) variable?

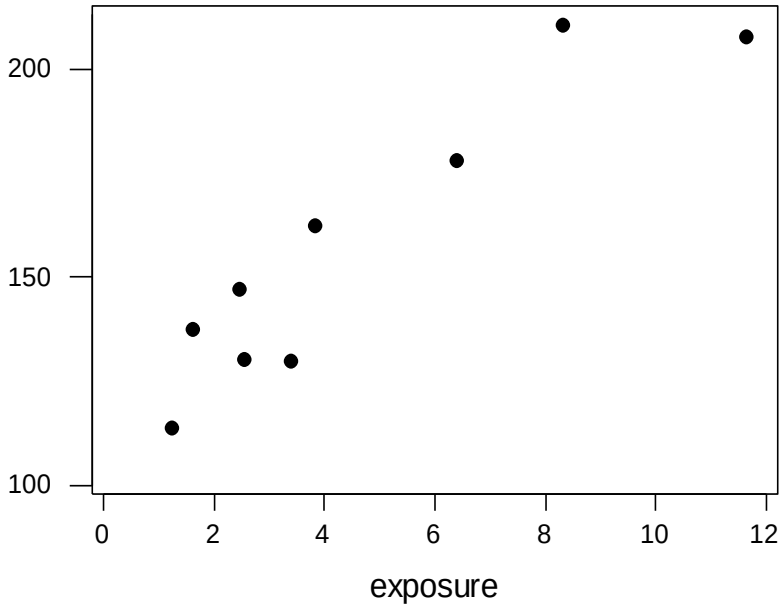
For instance, what if there is a genetic predisposition that causes people to both get lung cancer and become addicted to smoking, but the smoking itself doesn't CAUSE lung cancer?

Smoking Skull by Van Gogh

We can evaluate the association using the following criteria:

1) The association is strong.
2) The association is consistent.
3) Higher doses are associated with stronger responses.
4) The alleged cause precedes the effect.
5) The alleged cause is plausible.

Radioactive Waste and the Columbia River



- the association is strong, r = 0.926
- higher doses and stronger responses, b > 0
- but … what about confounding or lurking variables?

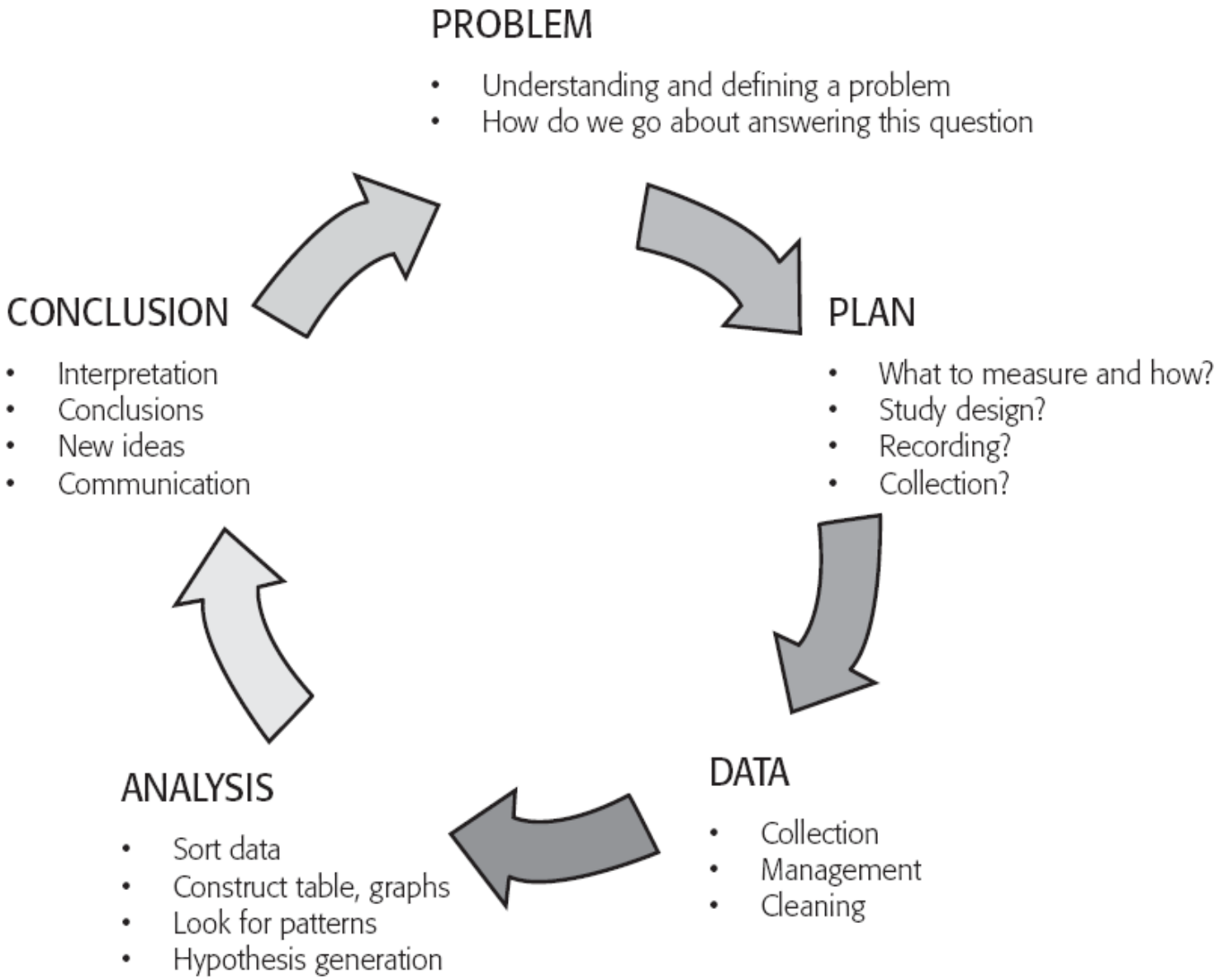
Cautions with regression

- Do not use a regression on inappropriate data.
  - Pattern in the residuals
  - Presence of large outliers
  - Clumped data falsely appearing linear

  } *Use residual plots for help.*
- Recognize when the correlation/regression is performed on averages.
- A relationship, however strong, does not itself imply causation.
- Beware of lurking variables.
- Avoid extrapolating (going beyond interpolation).

Putting Some Concepts Together

scatterplotting, time series plotting and linear regression


PROBLEM
• Understanding and defining a problem
• How do we go about answering this question
PLAN
• What to measure and how?
• Study design?
• Recording?
• Collection?
DATA
• Collection
• Management
• Cleaning
ANALYSIS
• Sort data
• Construct table, graphs
• Look for patterns
• Hypothesis generation
CONCLUSION
• Interpretation
• Conclusions
• New ideas
• Communication

Problem

- What impact did the events of September 11, 2001 have on travel by Americans?

Plan

- Get some measure of traffic flow out of the United States
  - use a measure suggested by Statistics Canada in their case study at: http://www.statcan.ca/english/edu/power/ch5/case_study/border.htm
    - border crossings of automobiles at the Canada/U.S. border
      - U.S. automobiles travelling into Canada at Fort Erie, Ontario

Consequences of the Study Plan

Pros

- data have already been collected, cleaned and made publicly available
- data source is reliable

Cons

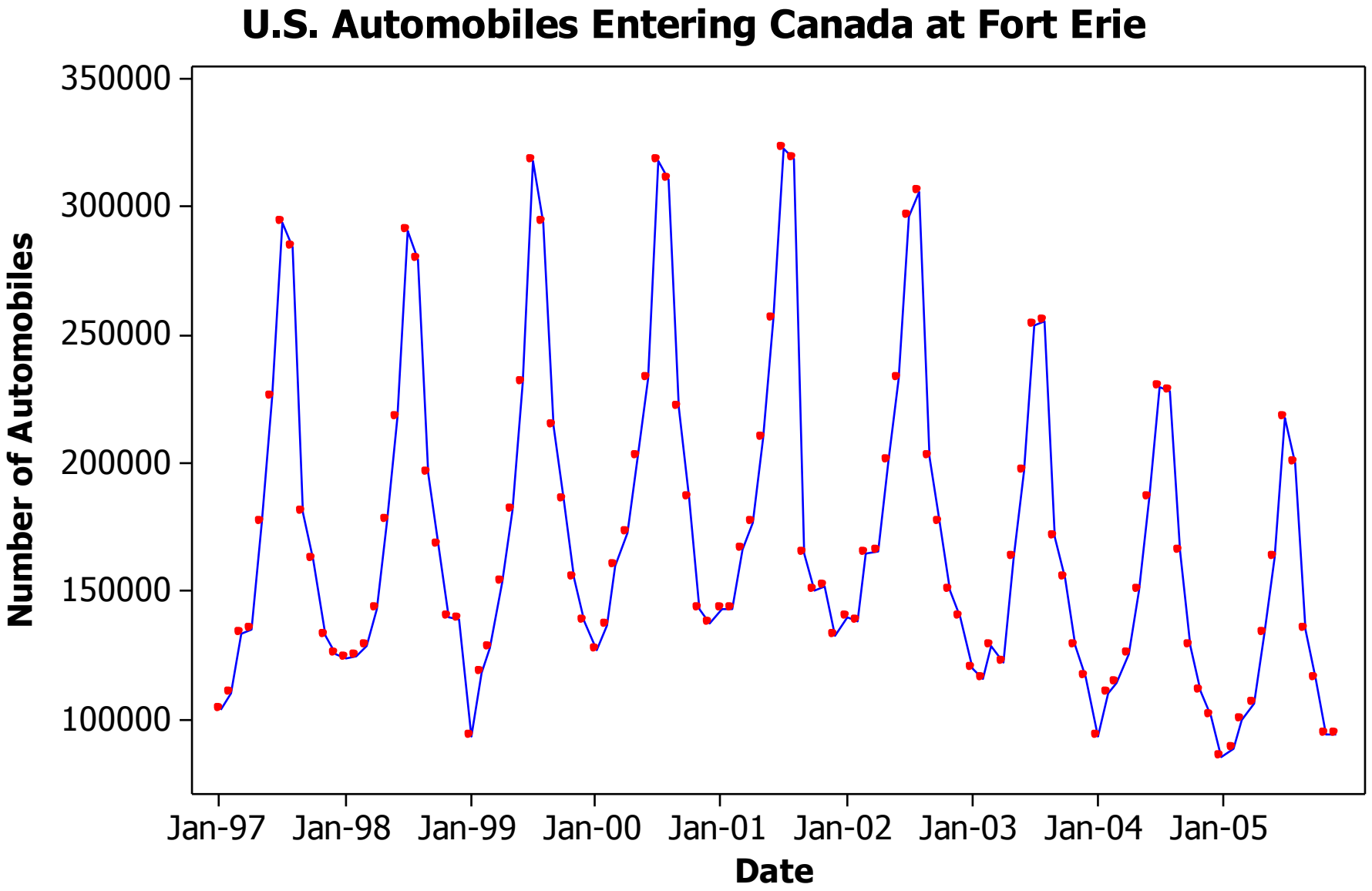
- results cannot be easily generalized (lack of a random sample)
  - to all U.S. travel to Canada
  - to all international travel done by people in the U.S.
- since this is an observational study, confounding or lurking variables may be present

Data

- data were taken from the CANSIM database (CANadian Socio Economic Information Management System) constructed and maintained by Statistics Canada
- data were chosen four years prior to 2001 and four years after 2001 to get a data series of reasonable length (January 1997 to December 2005) that straddled the key date of September 11, 2001.

Data: Number of U.S. Automobiles Crossing at Fort Erie

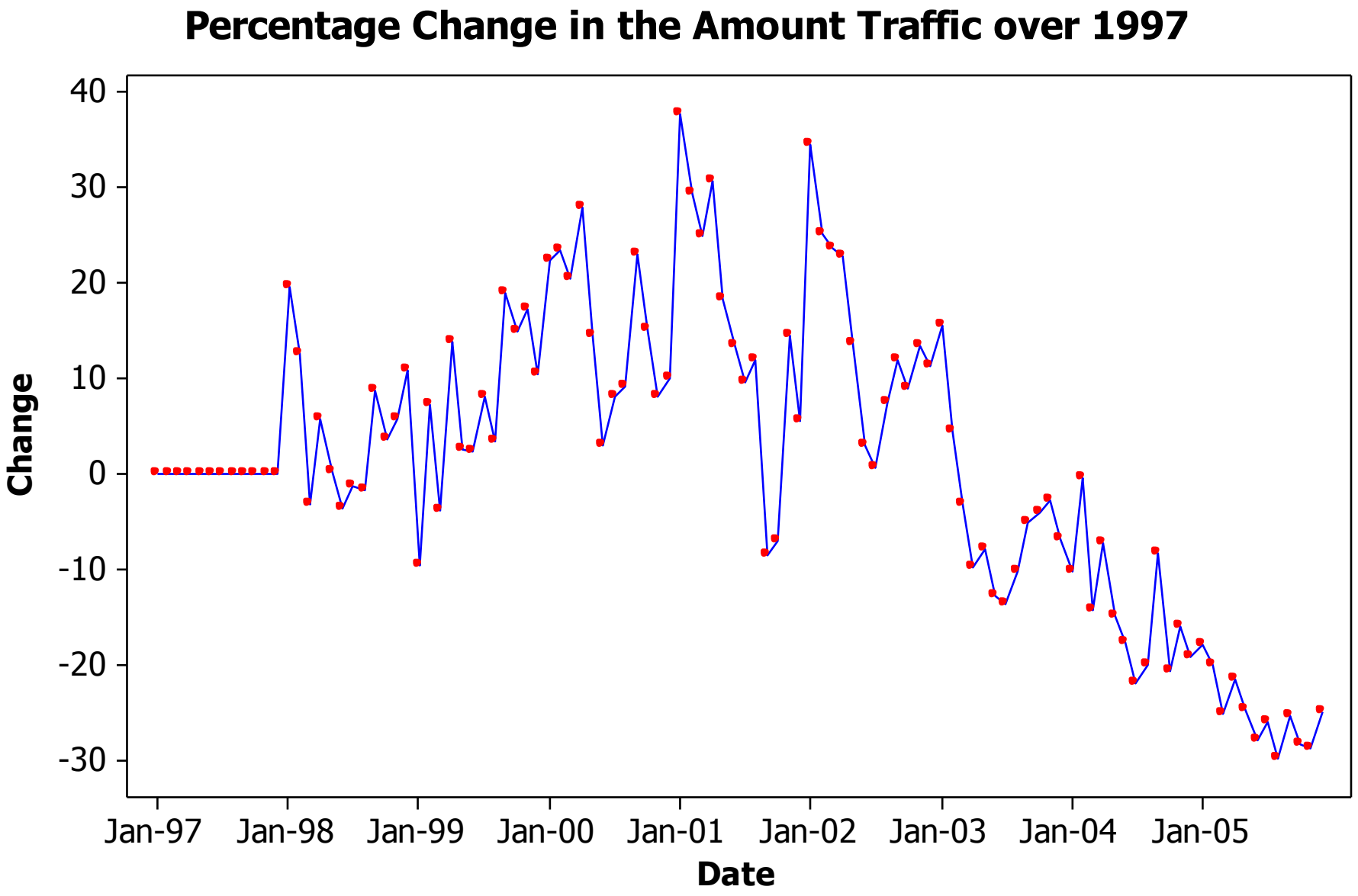
Month	1997	1998	1999	2000	2001	2002	2003	2004	2005
Jan	103732	123965	93733	126847	142810	139513	119912	93122	85123
Feb	110514	124507	118572	136482	142956	138388	115461	110100	88430
Mar	133146	128926	128180	160265	166293	164629	128954	114308	99607
Apr	135054	142931	153749	172661	176576	165711	121889	125280	105931
May	176897	177365	181447	202330	209239	200893	163057	150569	133308
Jun	225913	217691	231319	232621	256372	232597	196880	185977	163143
Jul	293941	290221	317574	317922	322337	295919	253982	229543	217497
Aug	284376	279733	293914	310244	318475	305683	255474	227685	199675
Sep	180408	196102	214658	221993	164920	202004	171394	165530	134899
Oct	161902	167618	185887	186521	150497	176388	155208	128553	116067
Nov	132404	139906	155096	143223	151545	150272	128759	111319	94281
Dec	125286	138854	138276	137835	132355	139550	116733	101307	94197


U.S. Automobiles Entering Canada at Fort Erie
Number of Automobiles
350000
300000
250000
200000
150000
100000
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Date

Observations

- the data have obvious seasonal variation
- de-seasonalize the data by looking at the percentage change in travel compared to 1997
- for month *x* (*x* = Jan, Feb, …, Dec) in year *y* (*y* = 1998, 1999, …, 2005) calculate

$$100\left(\frac{\text{Traffic in month } x \text{ of year } y}{\text{Traffic in month } x \text{ of } 1997} - 1\right)$$

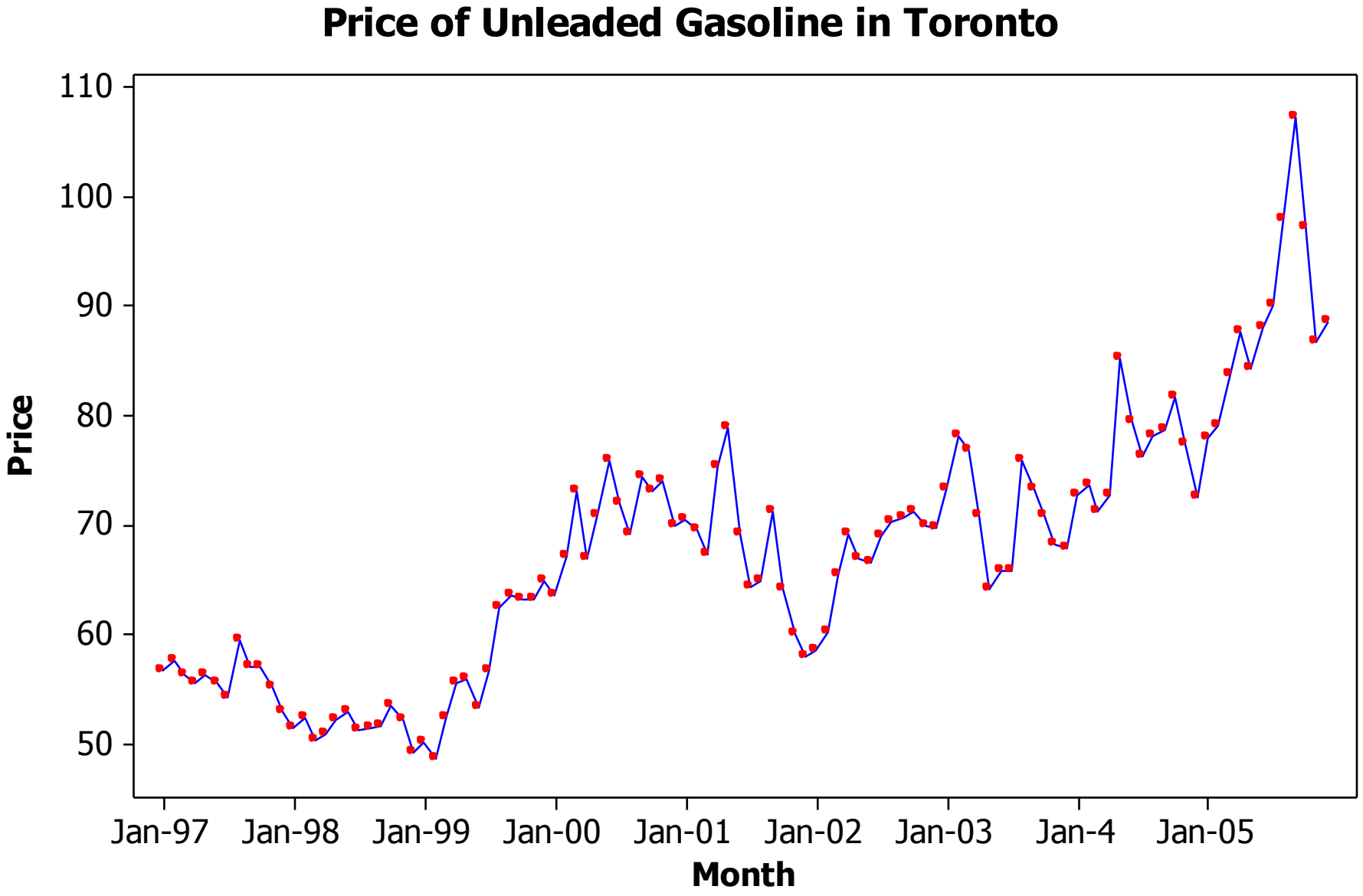

Percentage Change in the Amount Traffic over 1997
Change
40
30
20
10
0
-10
-20
-30
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Date

Observations

- general increase in automobile traffic prior to the latter part of 2001.
- large drop in the last few months of 2001 followed by a rebound
- general decrease in traffic after the rebound in 2002
  - Are there other explanations for the general decline in 2002 and later?
    - cost of gasoline
    - change in the exchange rate of the Canadian dollar

Data: Price of Unleaded Gasoline in Toronto at Self-Serve Pumps

Month	1997	1998	1999	2000	2001	2002	2003	2004	2005
Jan	56.7	51.4	50.2	63.5	70.4	58.6	73.3	72.8	78.0
Feb	57.5	52.4	48.6	67.1	69.6	60.2	78.2	73.7	79.0
Mar	56.3	50.3	52.4	73.1	67.3	65.5	76.8	71.2	83.8
Apr	55.6	50.9	55.5	66.9	75.4	69.1	70.9	72.8	87.6
May	56.2	52.2	55.9	70.8	78.8	66.9	64.2	85.2	84.2
Jun	55.6	52.9	53.3	75.8	69.2	66.5	65.8	79.5	88.1
Jul	54.3	51.2	56.6	72.0	64.4	69.0	65.8	76.3	90.1
Aug	59.4	51.4	62.4	69.2	64.8	70.3	75.9	78.2	97.9
Sep	57.1	51.7	63.6	74.4	71.2	70.7	73.2	78.7	107.2
Oct	57.0	53.4	63.2	73.0	64.2	71.2	70.8	81.6	97.2
Nov	55.1	52.2	63.1	74.1	60.1	70.0	68.3	77.3	86.7
Dec	52.9	49.2	64.9	70.0	57.9	69.7	67.8	72.5	88.5


Price of Unleaded Gasoline in Toronto
110
100
90
80
70
60
50
Price
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-4
Jan-05
Month

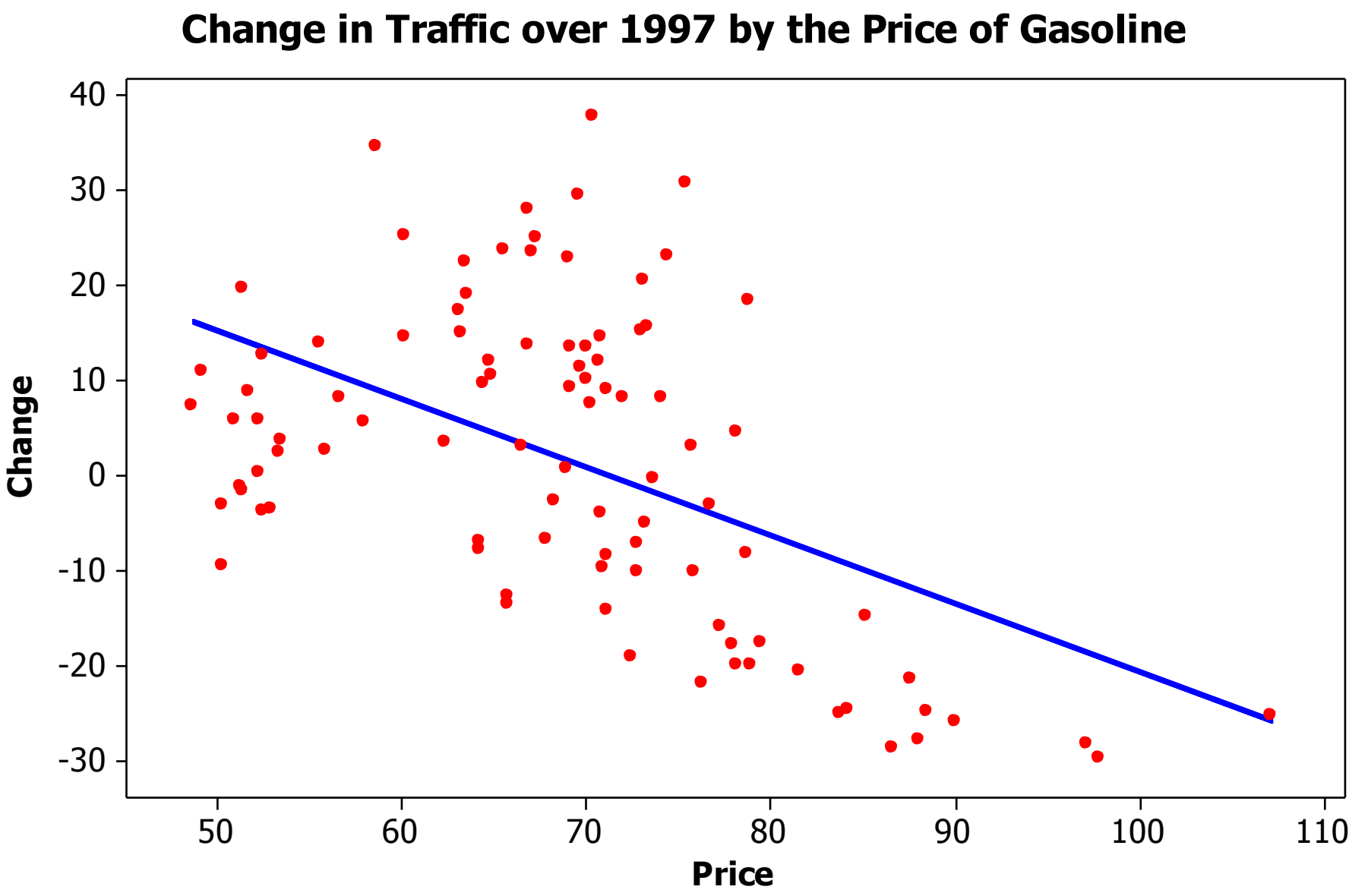
Observations and Procedure

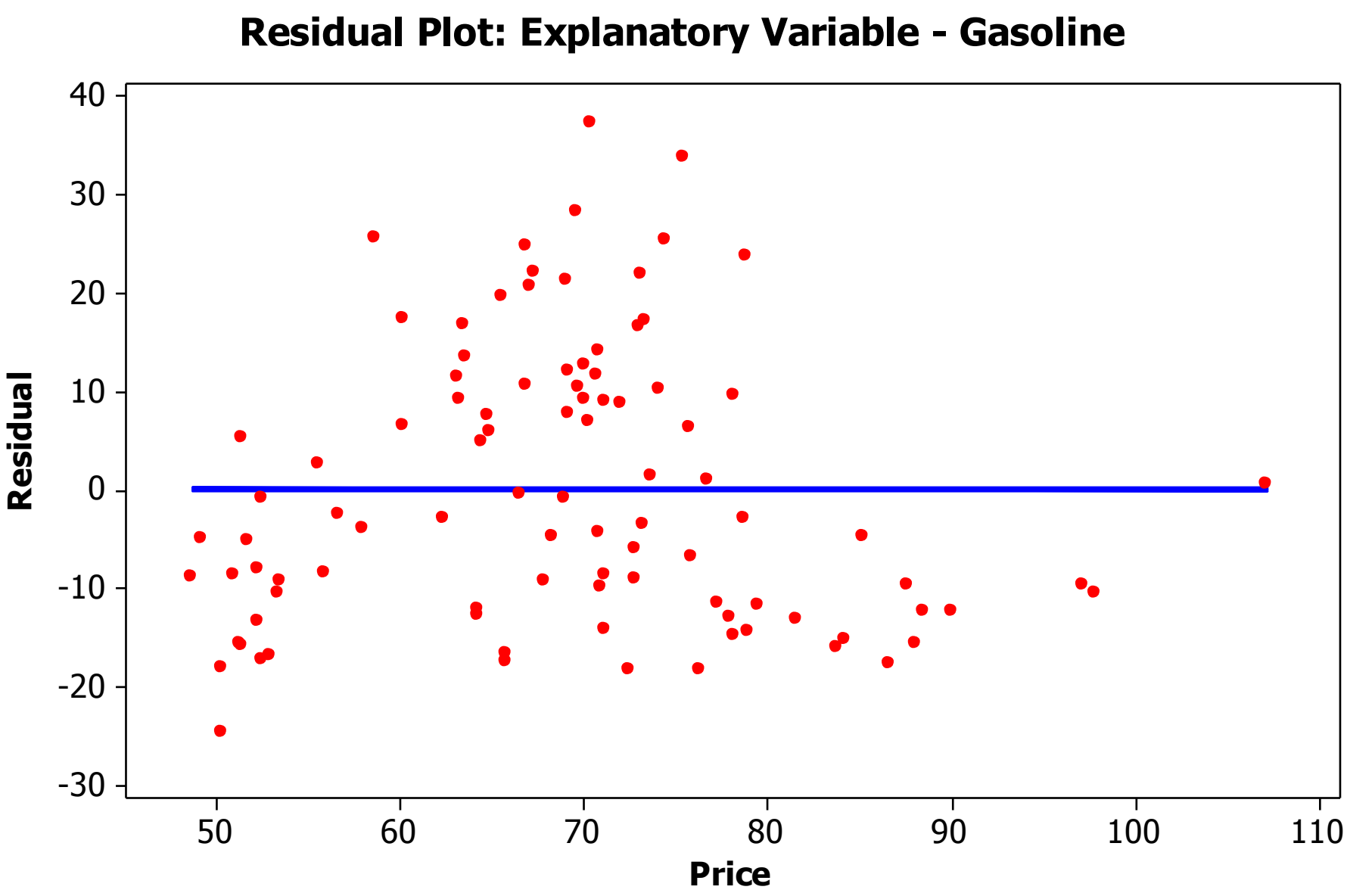
Observation

- if there is any seasonal variation in gasoline prices, it is not as pronounced as the traffic flow at the border

Procedure

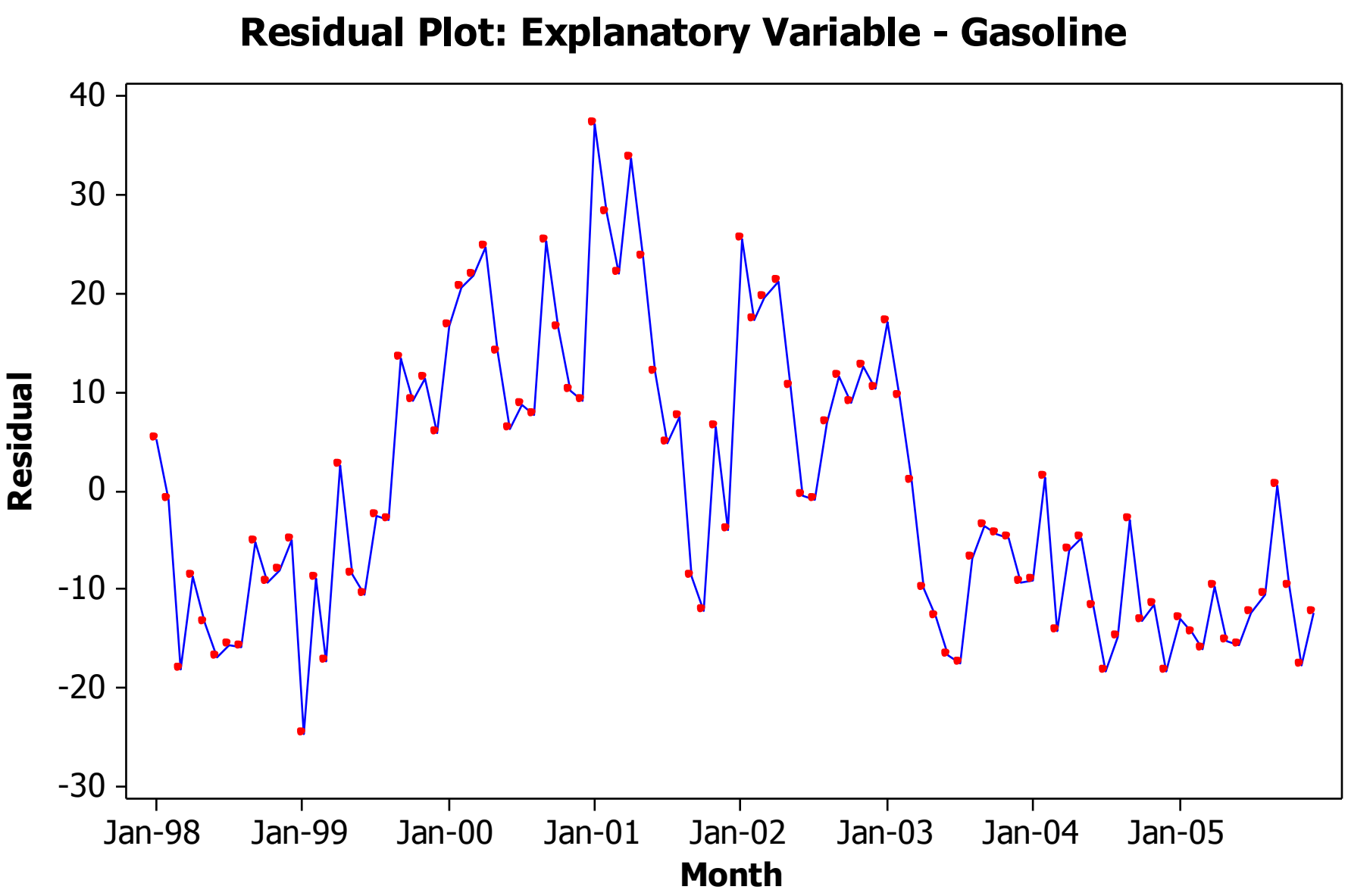
- plot the de-seasonalized traffic flow against gasoline prices (ignore 1997 the base year) and fit a regression line


Change in Traffic over 1997 by the Price of Gasoline
Change
40
30
20
10
0
-10
-20
-30
50
60
70
80
90
100
110
Price


Residual Plot: Explanatory Variable - Gasoline
Residual
40
30
20
10
0
-10
-20
-30
50
60
70
80
90
100
110
Price

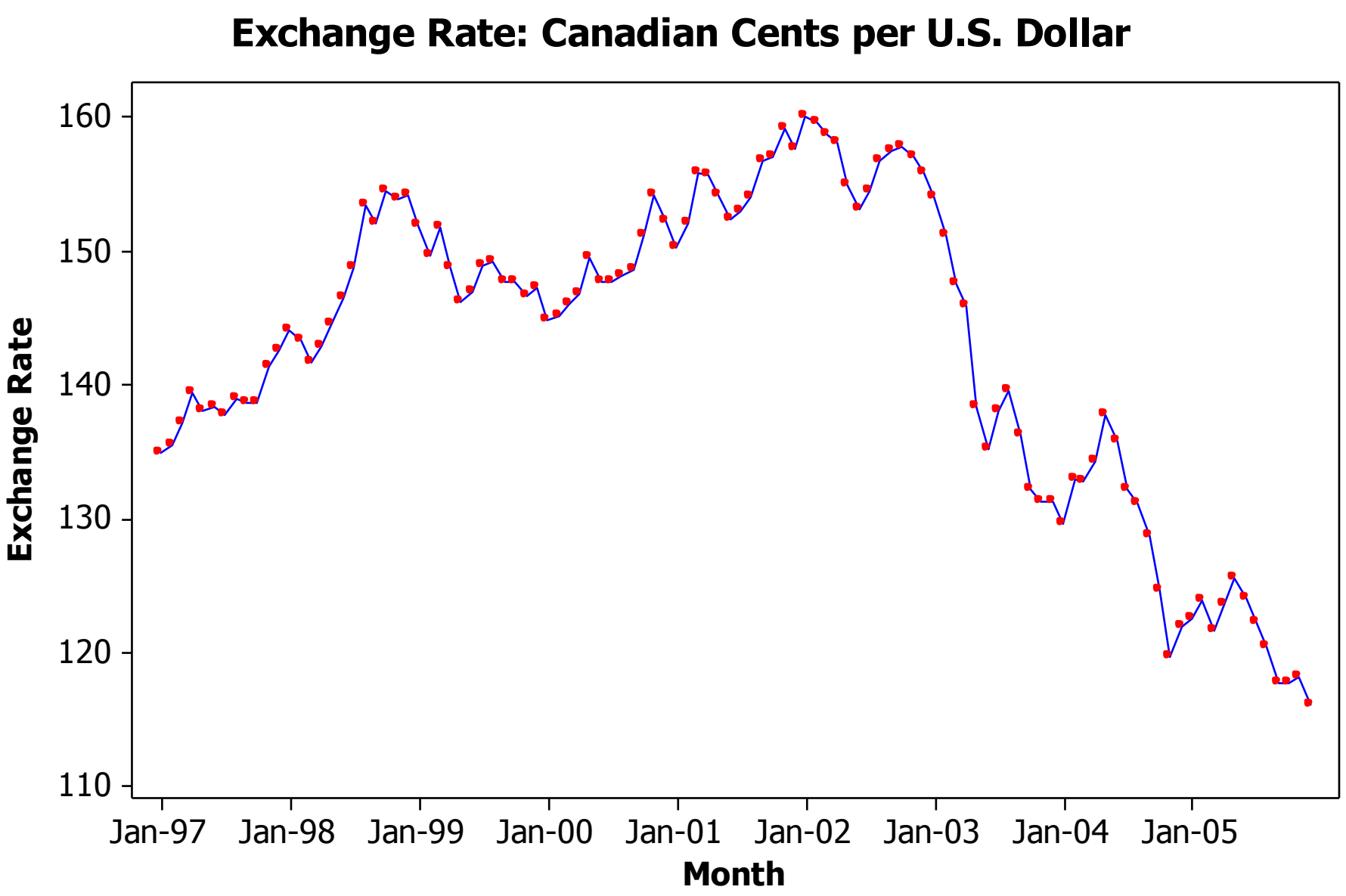
Observations

- a straight line model does not fit the data well
  - when the price is very low or very high the residuals tend to be negative in value
  - when the price is mid-range the residual are positive and negative, but positive residual tend to be larger in absolute value than the negative ones
- using a straight line model to remove the effect of gasoline prices on travel may not be a good idea


Residual Plot: Explanatory Variable - Gasoline
Residual
40
30
20
10
0
-10
-20
-30
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Month

Data: Exchange Rate – Number of Canadian Cents to Buy One U.S. Dollar

Month	1997	1998	1999	2000	2001	2002	2003	2004	2005
Jan	134.86	144.08	151.92	144.89	150.32	160.03	154.1	129.60	122.53
Feb	135.52	143.40	149.73	145.11	152.18	159.58	151.24	132.90	123.97
Mar	137.18	141.63	151.75	146.06	155.85	158.70	147.59	132.84	121.61
Apr	139.40	142.98	148.74	146.84	155.75	158.14	145.85	134.25	123.60
May	138.05	144.50	146.20	149.55	154.15	154.97	138.45	137.83	125.55
Jun	138.40	146.53	146.91	147.68	152.44	153.17	135.23	135.77	124.02
Jul	137.71	148.76	148.88	147.79	153.04	154.59	138.15	132.19	122.27
Aug	139.04	153.53	149.23	148.24	154.02	156.79	139.56	131.18	120.40
Sep	138.69	152.13	147.68	148.62	156.77	157.58	136.32	128.78	117.76
Oct	138.67	154.50	147.73	151.23	157.12	157.78	132.18	124.69	117.76
Nov	141.33	153.94	146.75	154.22	159.24	157.14	131.26	119.61	118.11
Dec	142.67	154.22	147.33	152.24	157.75	155.93	131.28	121.91	116.10


Exchange Rate: Canadian Cents per U.S. Dollar
Exchange Rate
160
150
140
130
120
110
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Month

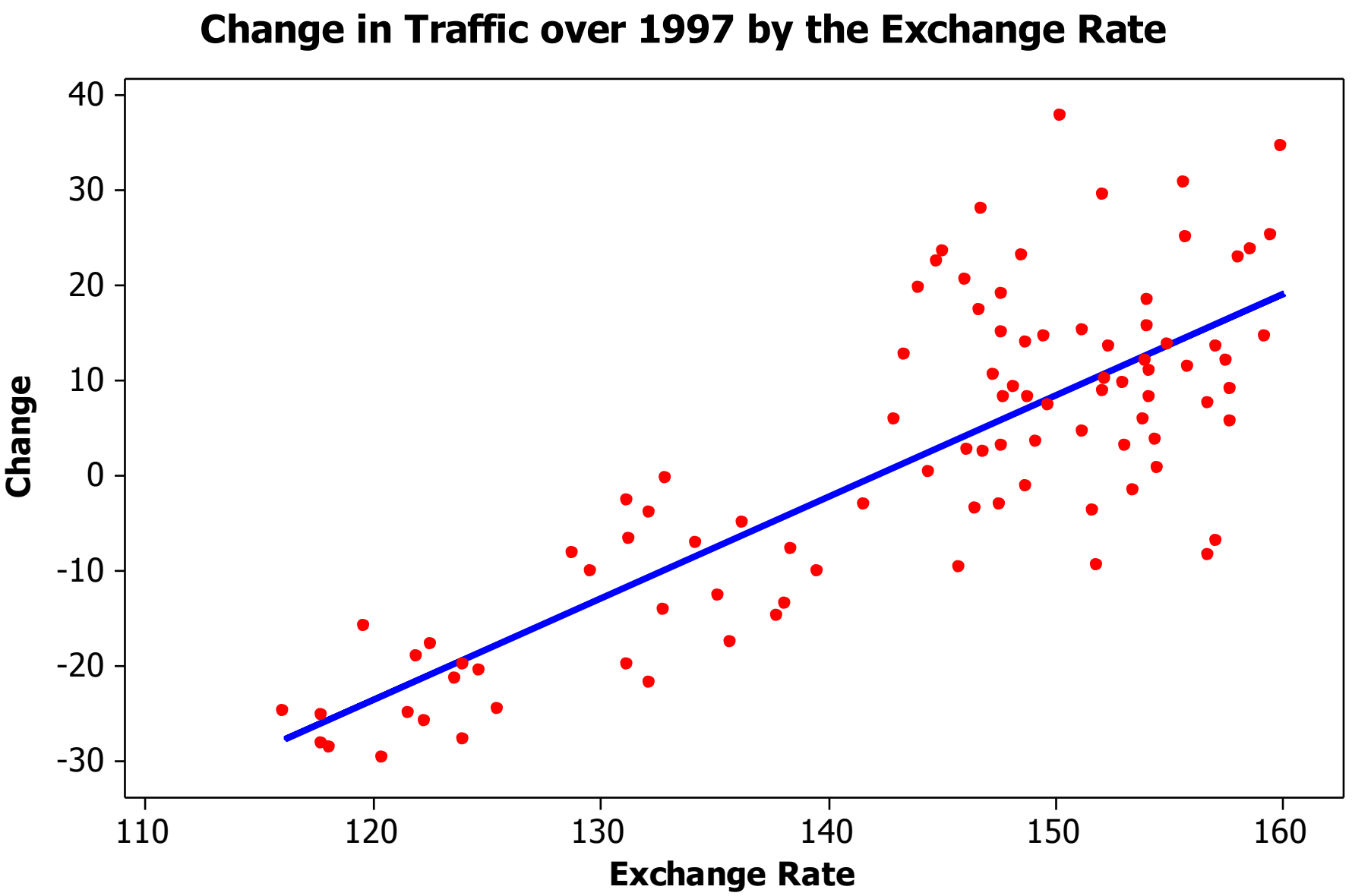
Observations and Procedure

Observation

- if there is any seasonal variation in the exchange rate, it is not as pronounced as the traffic flow at the border

Procedure

- plot the de-seasonalized traffic flow against the exchange rate (ignore 1997 the base year) and fit a regression line


Change in Traffic over 1997 by the Exchange Rate
Change
40
30
20
10
0
-10
-20
-30
110
120
130
140
150
160
Exchange Rate

Recall

$$\hat{y} = a + bx$$

$$b = r\frac{s_y}{s_x}$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{(n-1)s_x s_y}$$

$$a = \bar{y} - b\bar{x}$$

$$\bar{x} = \frac{134.86 + 135.52 + \cdots + 118.11 + 116.10}{108} = 143.02$$

$$\bar{y} = \frac{103732 + 110514 + \cdots + 94281 + 94197}{108} = 172239$$

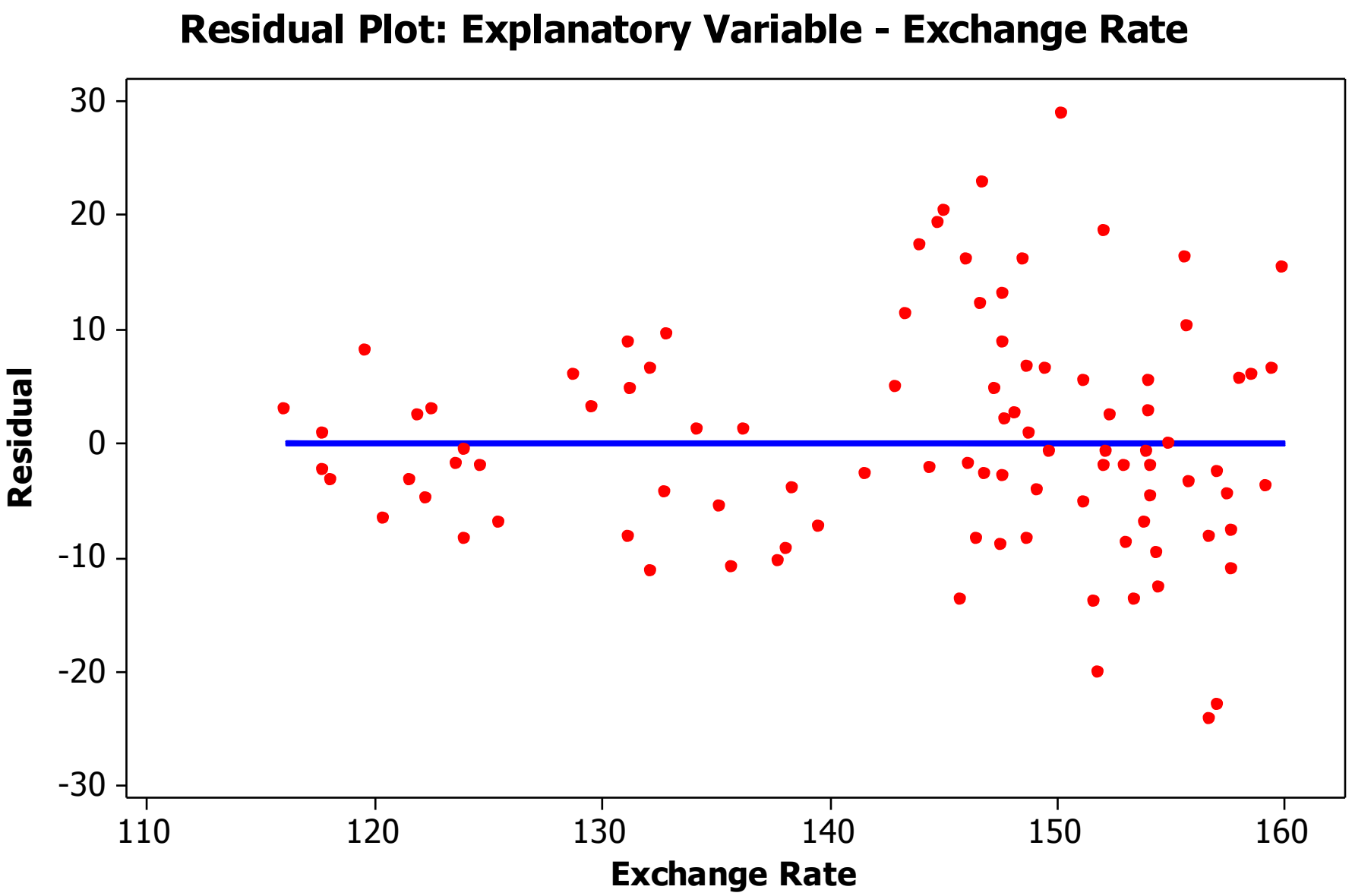
$$s_x = \sqrt{\frac{(134.86 - 143.02)^2 + (135.52 - 143.02)^2 + \cdots + (116.10 - 143.02)^2}{108 - 1}} = 11.80$$

$$s_y = \sqrt{\frac{(103732 - 172239)^2 + (110514 - 172239)^2 + \cdots + (94197 - 172239)^2}{108 - 1}} = 5871$$

$$r = \frac{(134.86 - 143.02)(103732 - 172239) + \cdots + (116.10 - 143.02)(94197 - 172239)}{(108 - 1)(11.80)(5871)} = 0.321$$

$$b = 0.321\frac{5871}{11.80} = 1662$$

$$a = 172239 - (1662)(11.80) = -65413$$


Residual Plot: Explanatory Variable - Exchange Rate
Residual
30
20
10
0
-10
-20
-30
110
120
130
140
150
160
Exchange Rate

Recall Residuals

$$y_i - \hat{y}_i$$

$$i = 1, \ldots, n$$

Observed (y_i)	x_i	Predicted ($\hat{y}_i$) $-65413+1662x_i$	Residual ($y_i - \hat{y}_i$)
103732	134.86	158675	–54943
110514	135.52	159771	–49257
⋮	⋮	⋮	⋮
94281	118.11	130842	–36561
94197	116.10	127502	–33305

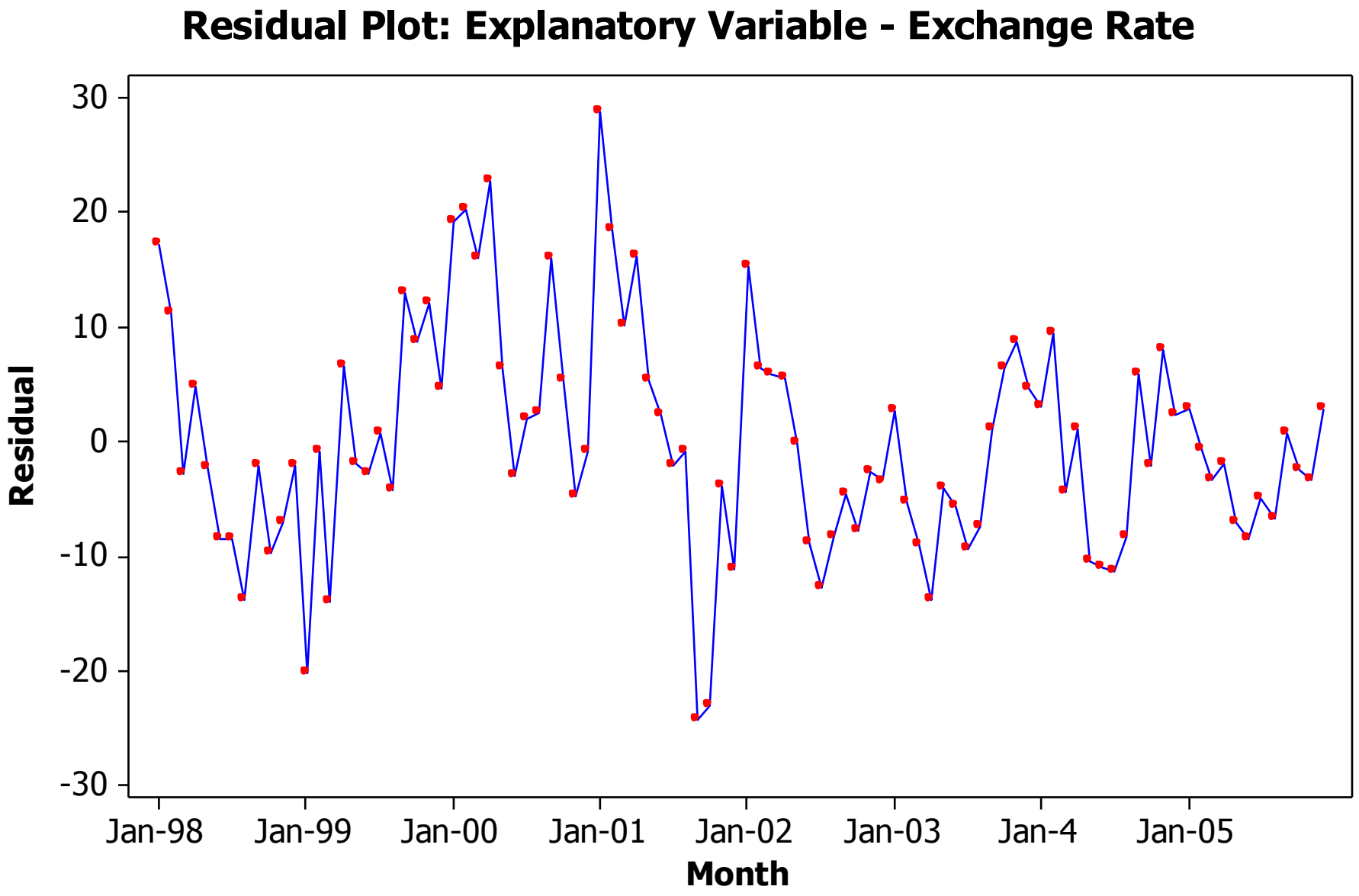
Observations and Procedure

Observations

- a straight line appears to fit the data
- variation around the line increases as the exchange rate increases (prediction of traffic flow at high exchange rates is more subject to error than at lower exchange rates)

Procedure

- plot the residuals (effect of the exchange rate has been removed) over time and look for the 9/11 effect


Residual Plot: Explanatory Variable - Exchange Rate
Residual
30
20
10
0
-10
-20
-30
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-4
Jan-05
Month

Conclusions

- the amount of U.S. automobile border crossings at Fort Erie
  - has monthly seasonal variation
  - is tied to the exchange rate (linearly)
- there was a drop in the amount of border traffic in a couple of months after 9/11
- further possible work could be done
  - for other automobile traffic across Canada (was there a less pronounced drop as we move west?)
  - for airline travel of U.S. passengers into Canada