AP Statistics

REGRESSION WISDOM
CHAP 8

Avoid Linear extrapolation … The turkey's first 1000 days are a seemingly unending succession of gradually improving circumstances confirmed by daily experience. What happens on Day 1001? Thanksgiving.

John E. Sener (1954 -)

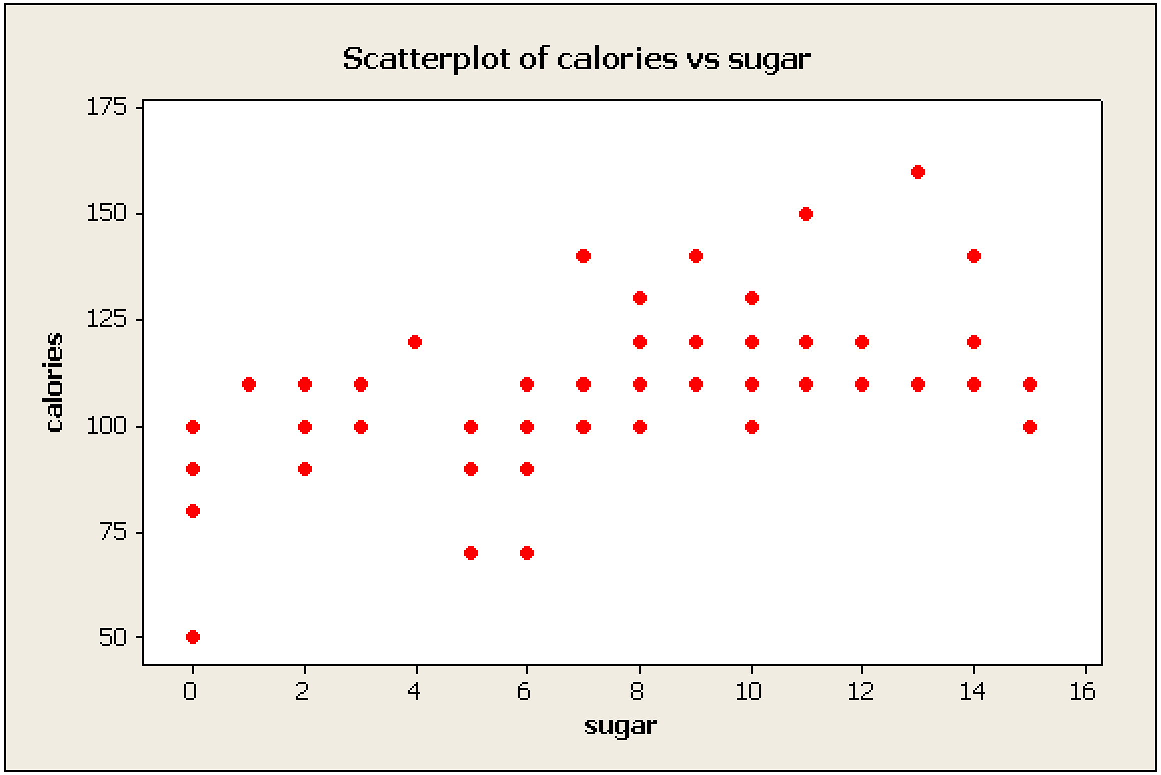
Look for Groups in the Residuals

Who – 77 breakfast cereals
What – sugar content (g) and calories

name	calories	sugar
100%_Bran	70	6
100%_Natural_Bran	120	8
All-Bran	70	5
All-Bran_with_Extra_Fiber	50	0
Almond_Delight	110	8
etc.		

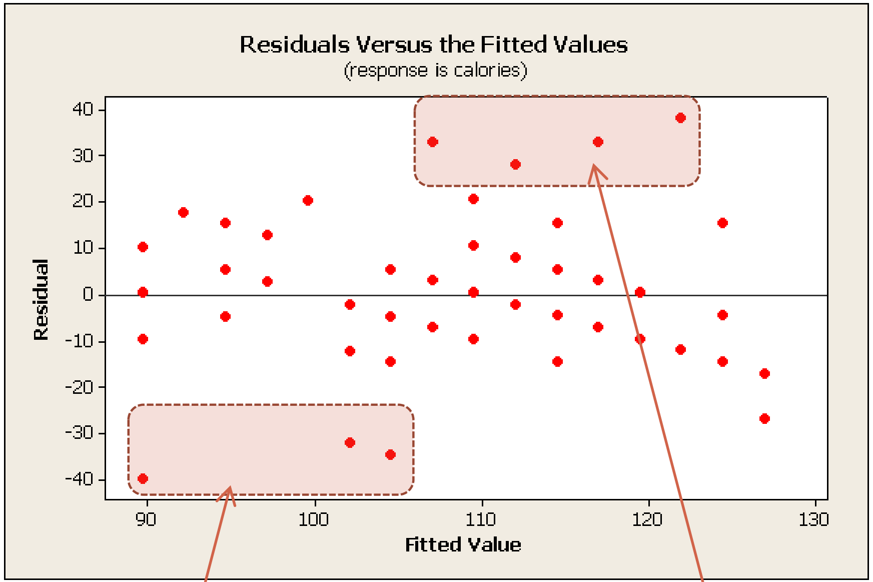
What is the association between the number of calories and the sugar content? Specifically, can we predict the number of calories from the amount of sugar?

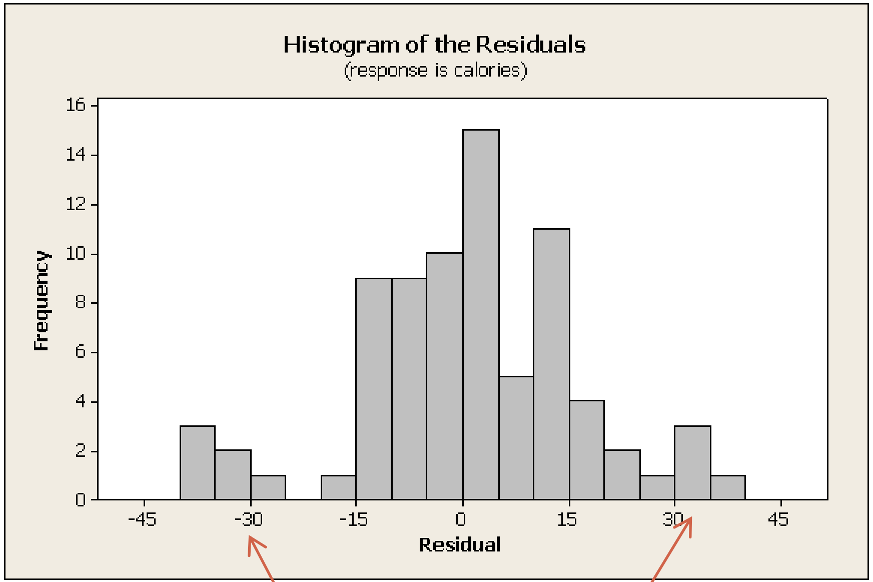
Look for Groups in the Residuals



There is a positive and moderately linear association between the amount of sugar and the number of calories in these breakfast cereals.

Look for Groups in the Residuals



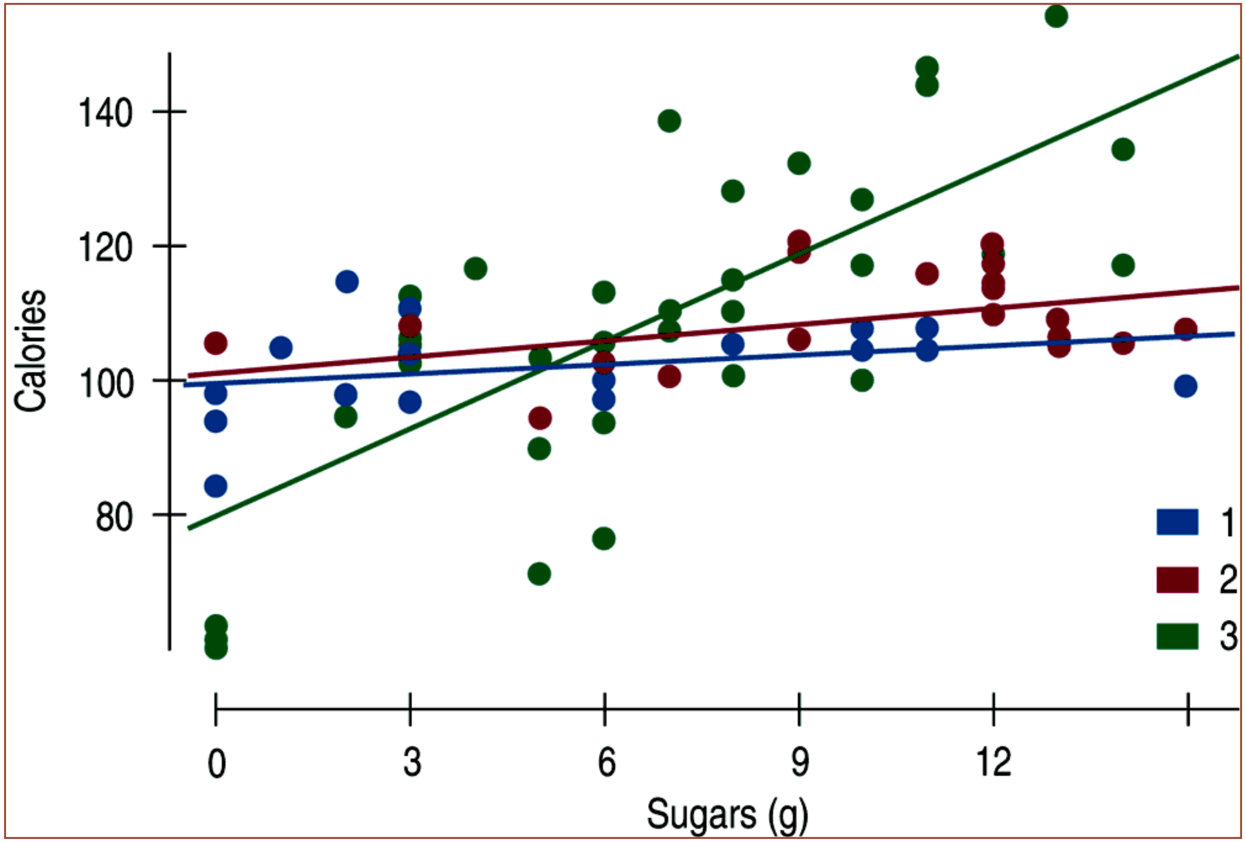


“Healthy cereals?”

multiple modes

Low calorie for
sugar content

Plot Subsets of Data



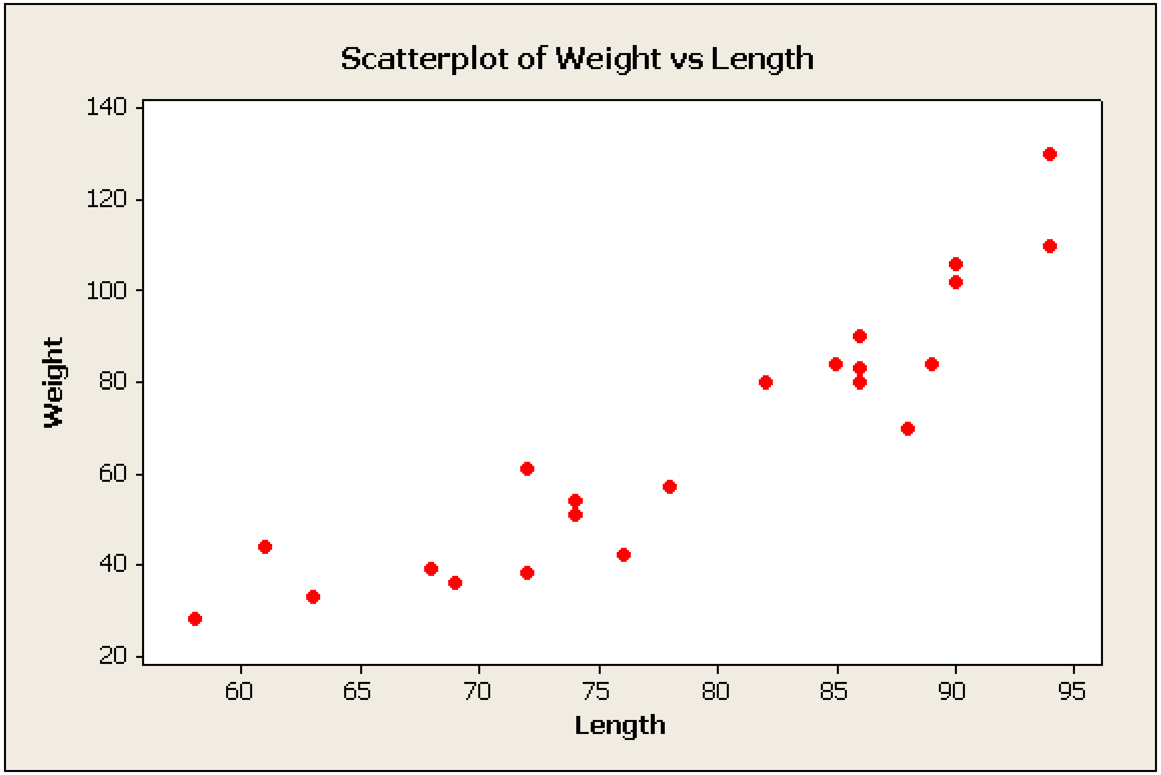
Separate the cereals into 3 groups:

1. bottom shelf
2. middle shelf
3. top shelf

Note the top shelf appears to be different from the other two.

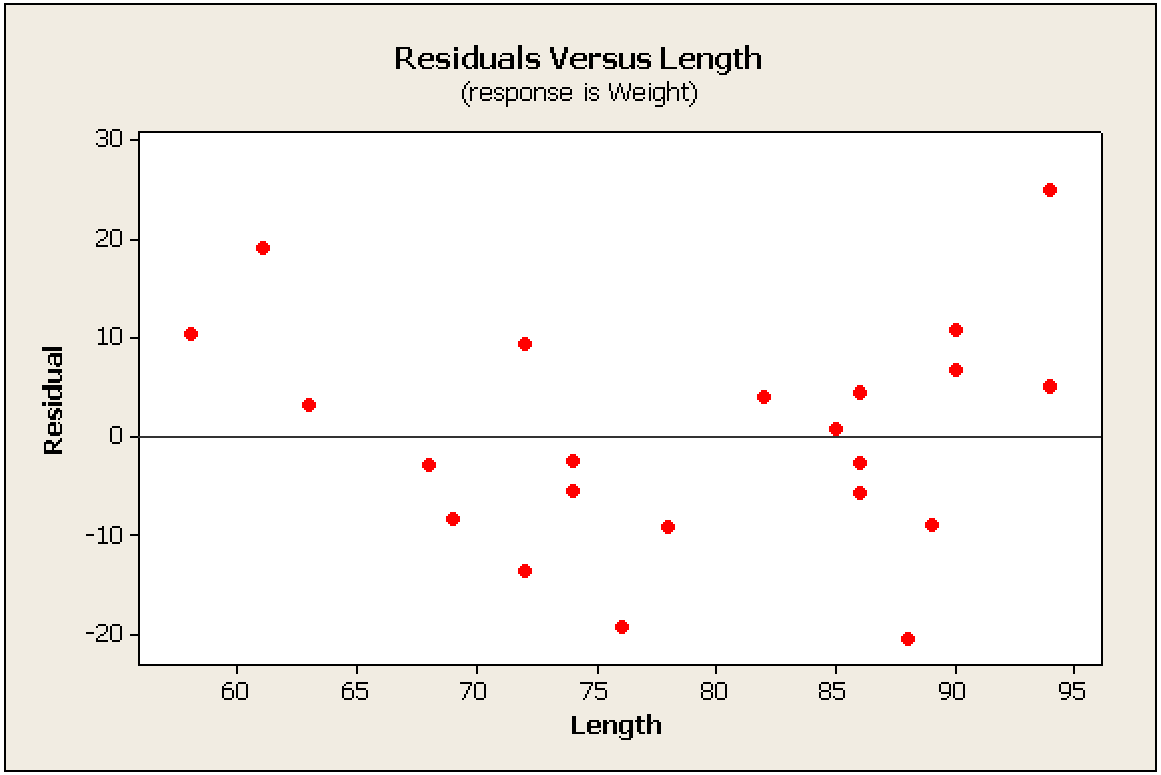
Look for Curves

Who – 22 alligators
What – length and weight



How would you describe the association between length and weight?

Look for Curves

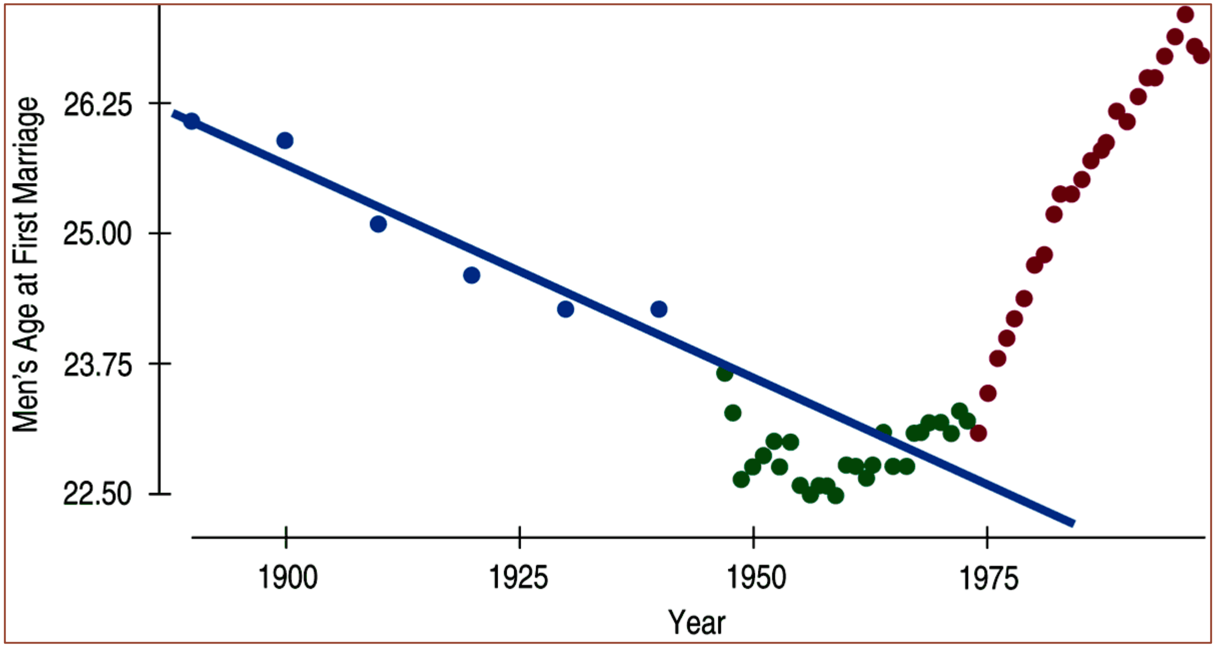


Notice the curve in the residuals.

The linear model did not fully capture the relationship between weight and length.

Extrapolation

Who – American men in 20th century
What – age at first marriage



A regression of mean age at first marriage for men vs. year fit to the first 4 decades of the 20th century does not hold for later years. You could split the data set into two parts and make one model for the first half of the century and another model for the second half.

Outliers, Leverage, and Influence

Outliers

- Data points that stand away from the others
- May have large residuals or high leverage

Leverage

- Data points whose *x*-values are far from the mean of *x*
- High-leverage points pull the line close to them, sometimes with large effect on slope
- Such points may have small residuals if they follow the pattern of the other data points

Influential Point

- Data point, when omitted, results in regression model with a very different slope

Outliers, Leverage, and Influence

Some fit the pattern and will have small residuals and have little effect on the model or R-squared:

With outlier (blue):

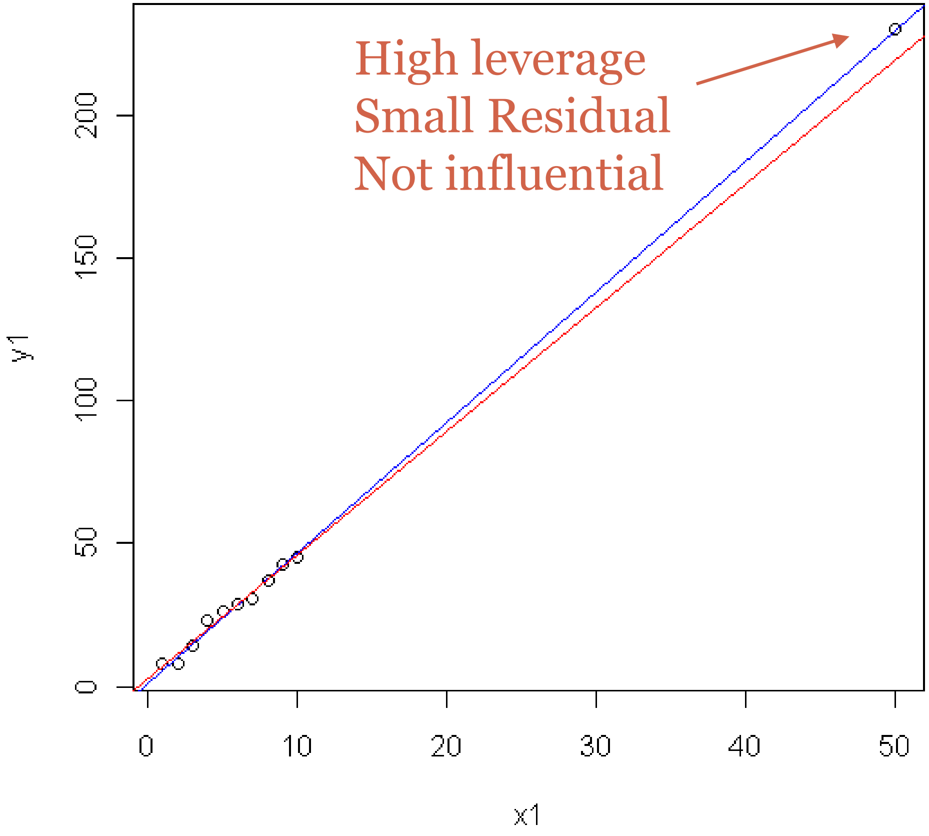
$$\hat{y} = 1.13 + 4.60x$$

$$R^2 = 99.9\%$$

Without outlier (red):

$$\hat{y} = 2.34 + 4.34x$$

$$R^2 = 97.9\%$$



High leverage
Small Residual
Not influential

Outliers, Leverage, and Influence

Some won't fit the pattern and will have large residuals and have little effect on the model, but a big effect on R-squared:

With outlier (blue):

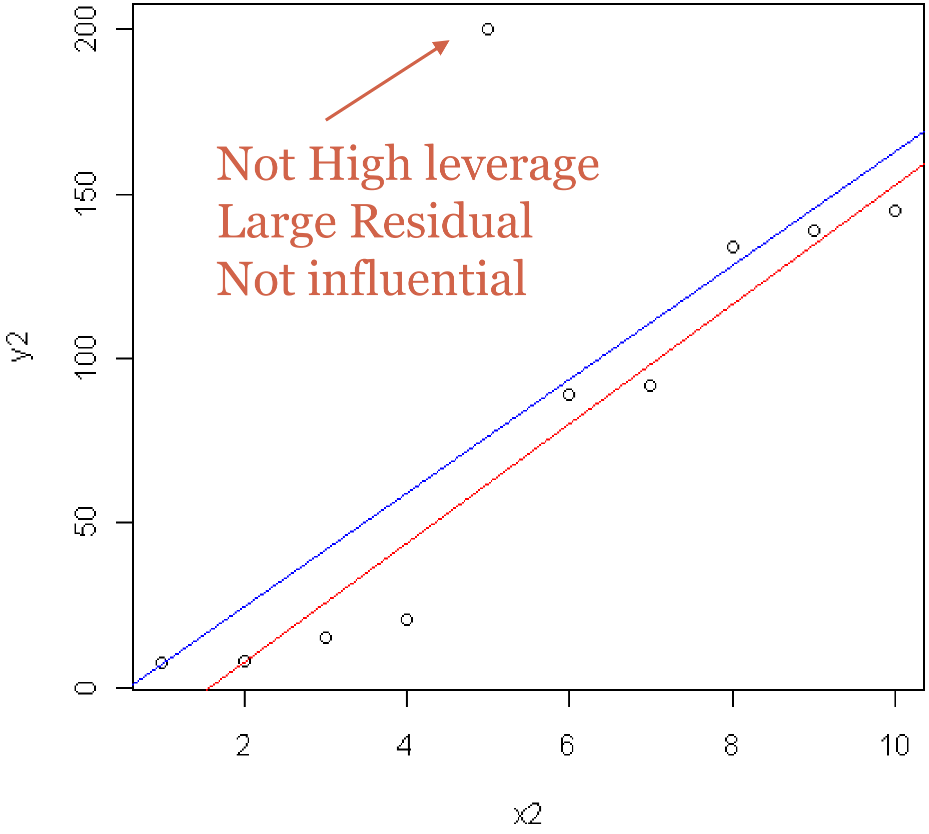
$$\hat{y} = -10.24 + 17.30x$$

$$R^2 = 57.1\%$$

Without outlier (red):

$$\hat{y} = -28.63 + 18.14x$$

$$R^2 = 94.8\%$$



Not High leverage
Large Residual
Not influential

Outliers, Leverage, and Influence

Some create the illusion of a stronger association than really makes sense. These “high leverage” points are considered to be “influential” because of the <u>extreme change in slope</u> when removed.

With outlier (blue):

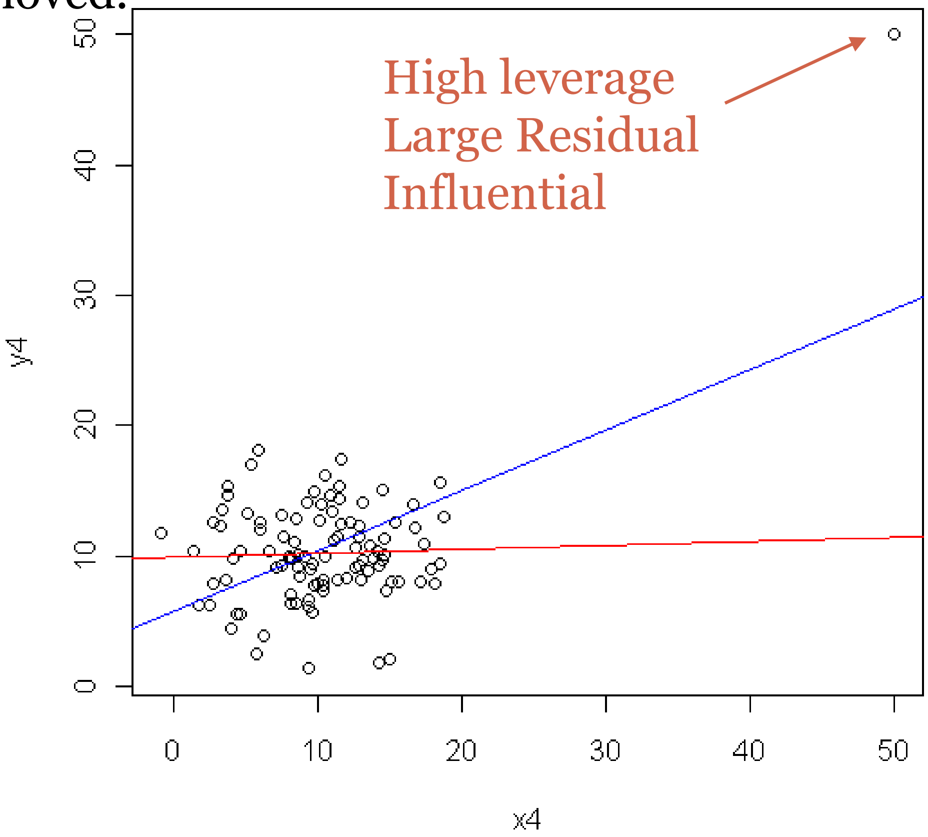
$$\hat{y} = 5.67 + 0.47x$$

$$R^2 = 27.8\%$$

Without outlier (red):

$$\hat{y} = 9.83 + 0.03x$$

$$R^2 = 0.2\%$$

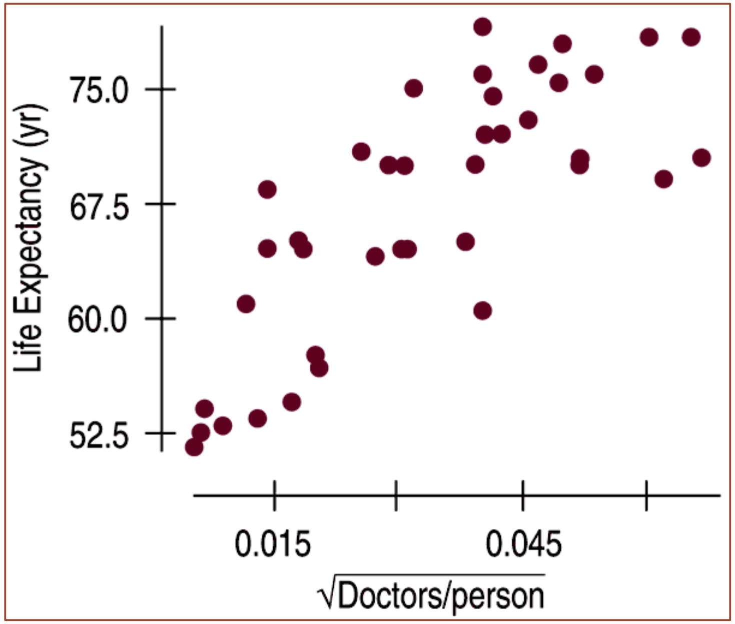


High leverage
Large Residual
Influential

Extraneous (Lurking) Variables

Extraneous variables are :

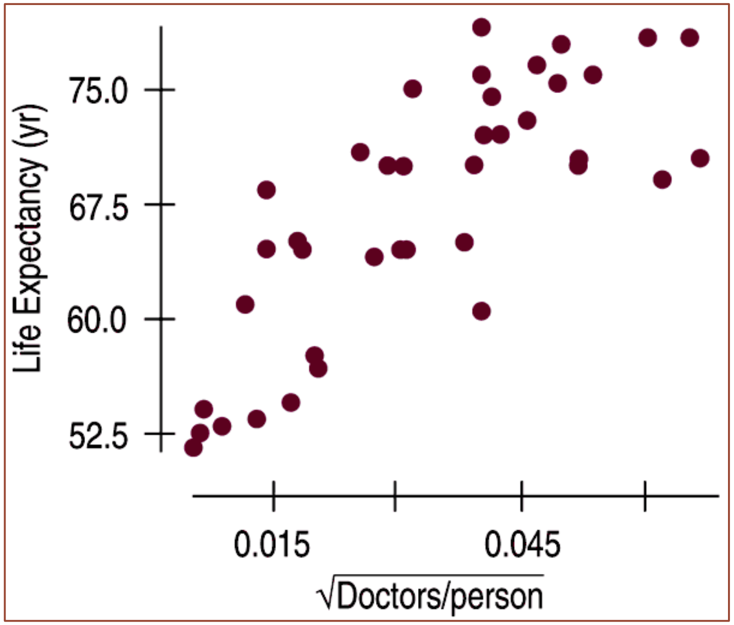
- variables other than the explanatory and response variables
- variables that may have an important influence on the association between the explanatory and response variables

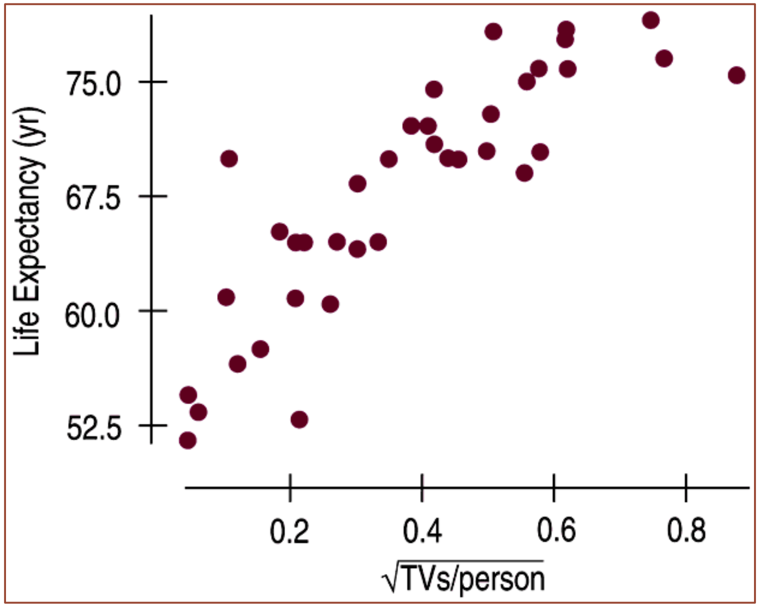


Average life expectancy is related to (the square root of) the number of doctors per person.

Proof that more doctors per person causes longer life expectancy?

Extraneous (Lurking) Variables





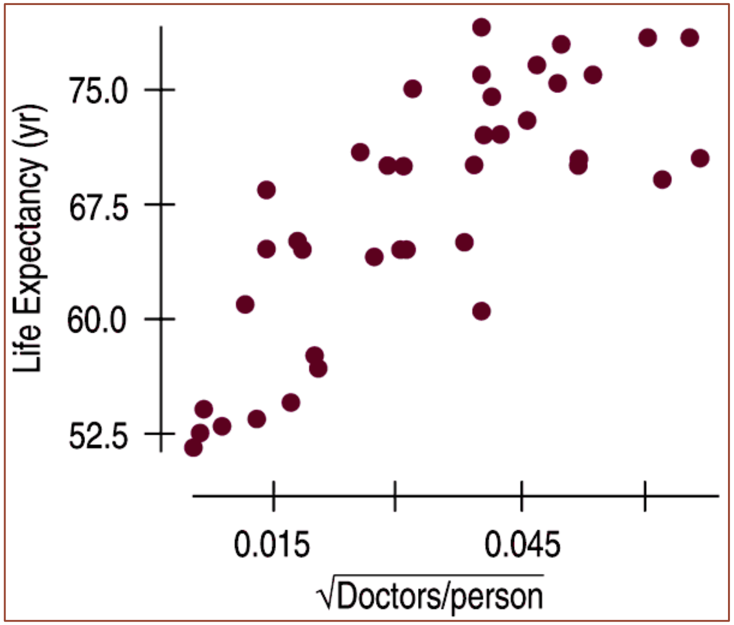
Average life expectancy is related to (the square root of) the number of doctors per person.

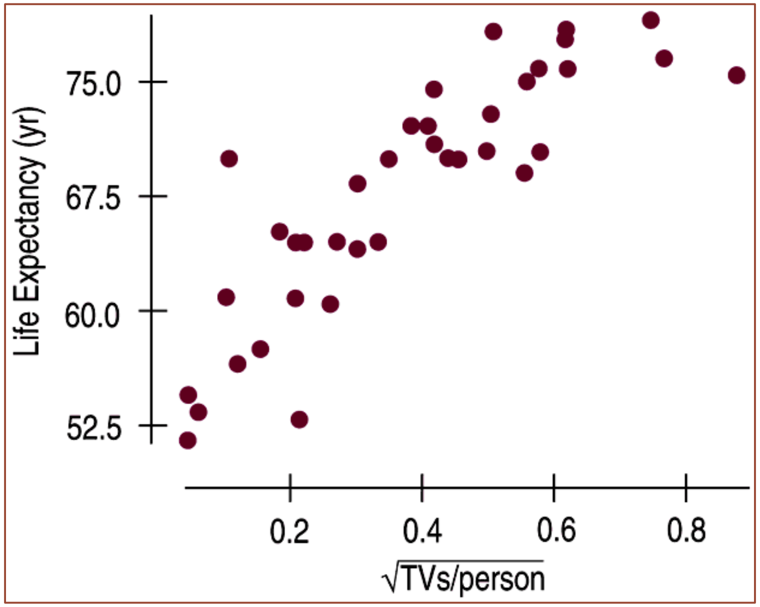
Proof that more doctors per person causes longer life expectancy?

Average life expectancy is related to (the square root of) the number of TVs per person.

Proof that more TVs per person causes longer life expectancy?

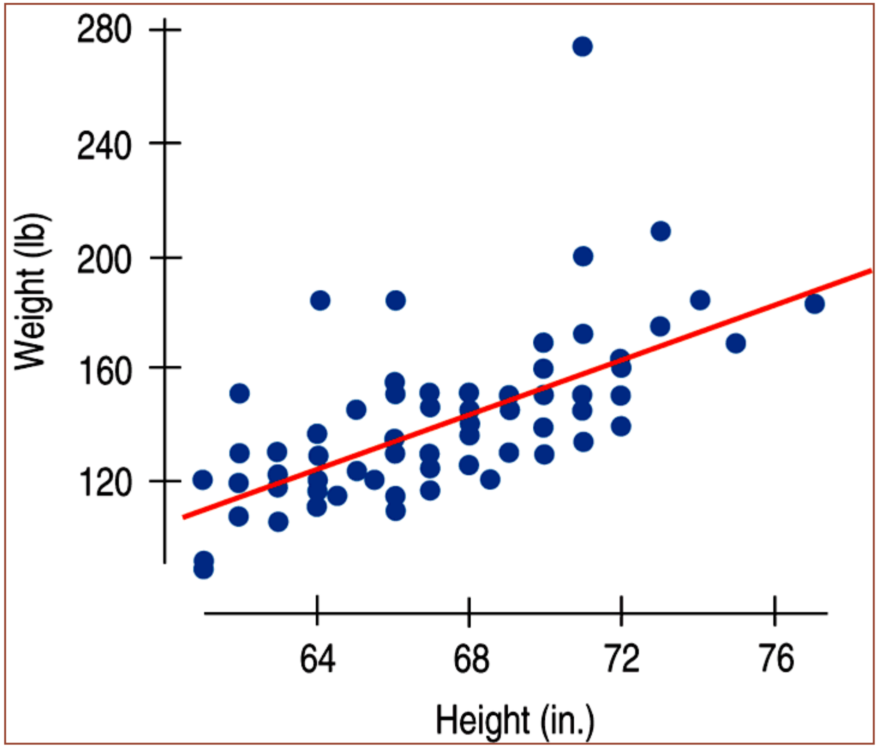
Extraneous (Lurking) Variables

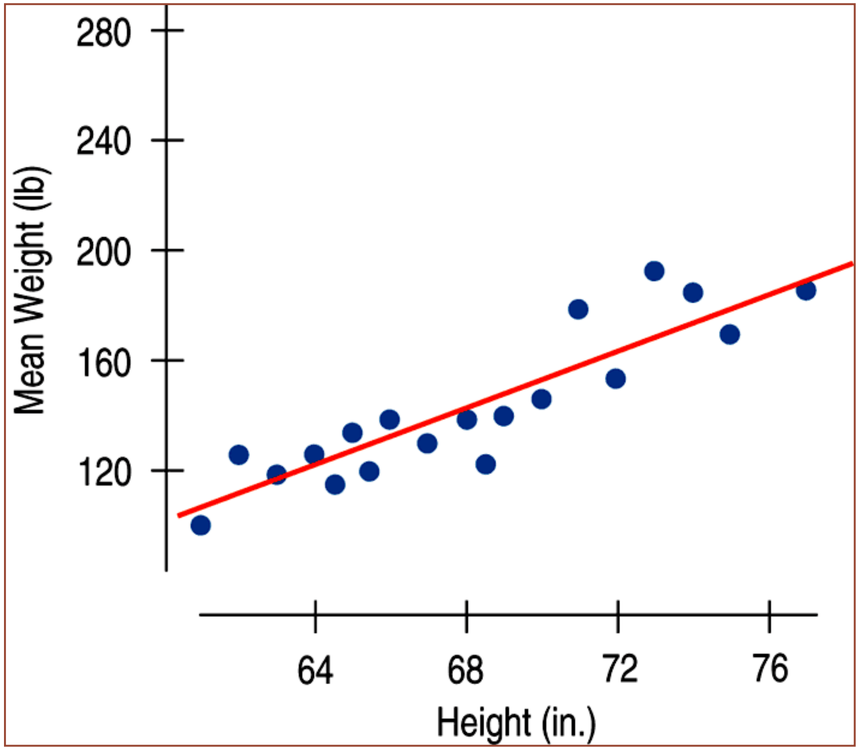




A better explanation is that wealth is a lurking variable. Countries with greater wealth are likely to have more doctors and other factors that contribute to longer life expectancy. More wealth also means greater access to luxury items such as TVs. A higher standard of living might be the cause of the observed relationships.

Summary Values are Less Variable





Weight vs. Height

Mean Weight vs. Height

The second plot suggests a stronger linear association than is really present.

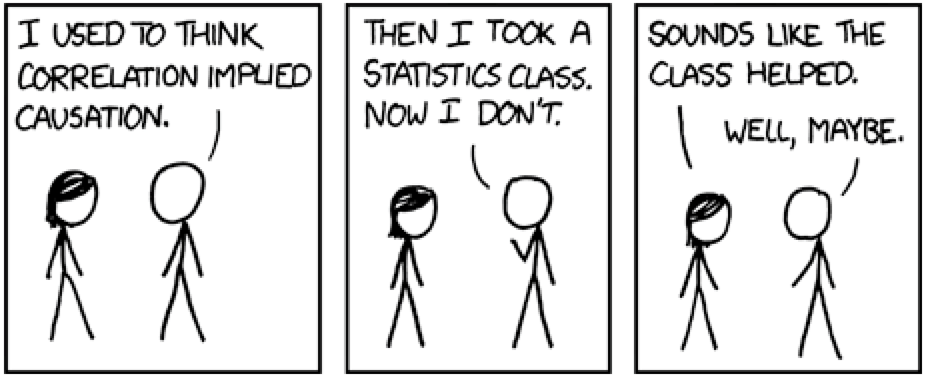
Regression Summary

- Make sure the relationship is straight – make both a scatterplot of the data and a scatterplot of the residuals.
- Look for subsets in your data. Fit different linear models to each group.
- Extrapolate with caution.
- Look for unusual points – correct or explain, if possible. Don't delete without due cause.
- Beware of influential points. Compare regressions with and without such points to see how influential they may be.
- Beware lurking variables. Association does not imply causation.
- Remember that summary data can inflate the strength of association.

Assignment

Read Chap 8

Chap 8 #1, 7-15 odd, 21, 25, 27, 29


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

www.xkcd.com