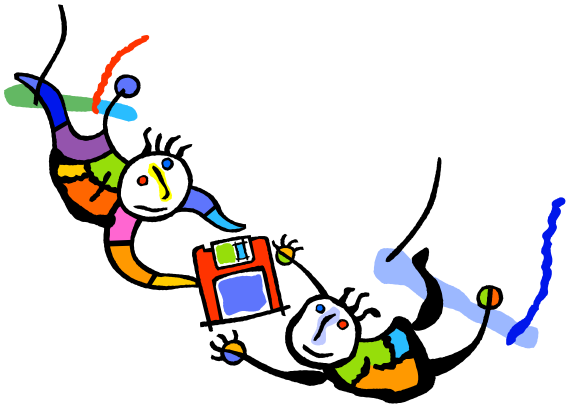


BIVARIATE DATA ANALYSIS

TWO QUESTIONS OF BIVARIATE DATA ANALYSIS

- What is the degree of linearity?

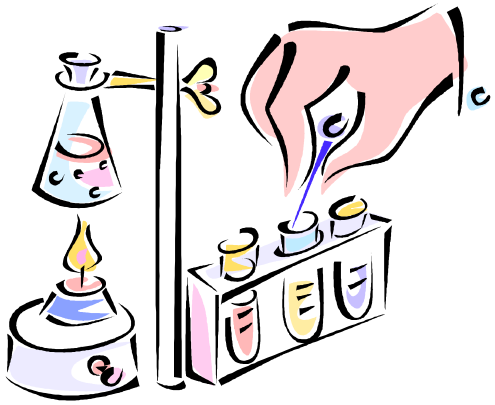
  is there a line?

- What is the degree of association?

  how strong is the line?

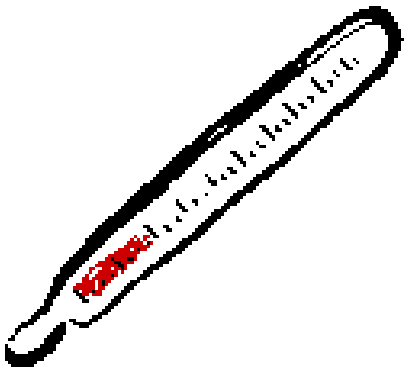
EXPLANATORY VARIABLE

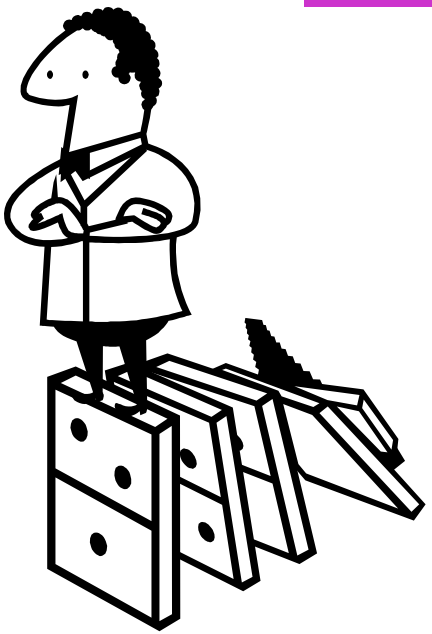
- Independent Variable
- What is the “x”?
- Usually what the researcher is trying to manipulate



RESPONSE VARIABLE

- Dependent Variable
- What is the “y”?
- What the researcher is trying to see an effect on



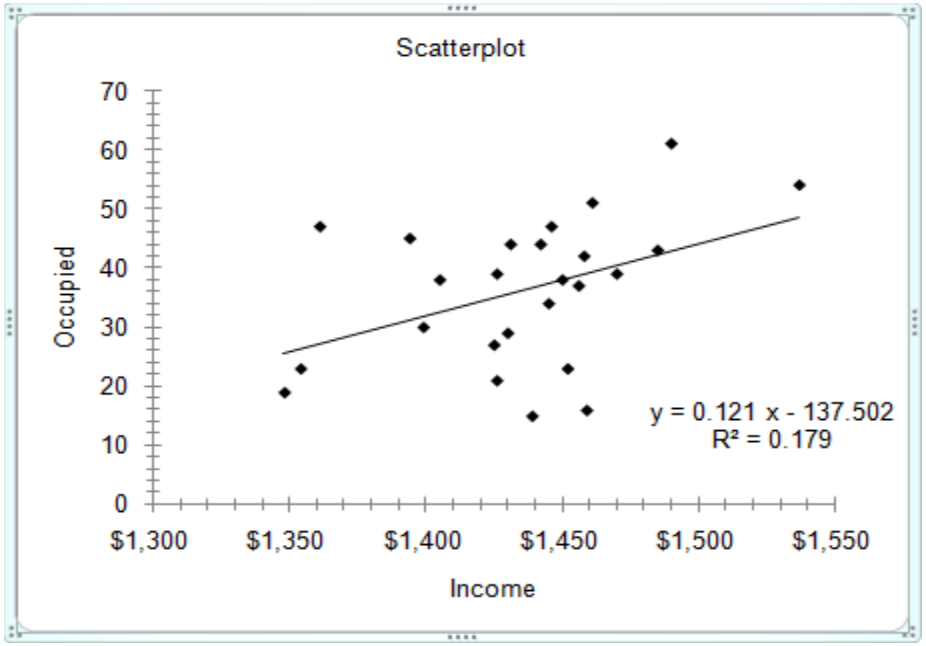


IF WE WANT TO PREDICT Y, WHAT COULD X BE?

1. Predict weight of human males
2. Predict Math SAT score
3. Predict college freshman GPA
4. Predict amount of growth of a plant
5. Predict blood alcohol content of driver
6. Predict cost of building a house

SCATTERPLOT

Definition: graph of two variables
with dot to represent each observation



"COMMENT ON THE SCATTERPLOT"

Shape: Does relationship look linear?

Outliers: Are there any unusual points?

Direction: Is linear relationship positive

or negative?

Strength: Is the line strong?

CORRELATION

Definition: Measures the strength of the linear relationship between variables

r is the symbol for correlation

r takes on values between -1 and 1

0 means no linear relationship

-1 and 1 mean perfect linear relationship

CORRELATION ≠ CAUSATION

Just because you have a high **r** value does not mean that x causes y…be careful!!

There could be a "Hidden variable" that actually is the cause.

MINUTES LATE TO WORK

1. Direct Causation: late to work and rain
2. Reverse Cause and Effect: late to work and poor relationship with boss
3. 3rd variable: late to work and rain and # of children at home
4. Coincidence: late to work and height

REGRESSION LINE

Linear model created by bivariate data set

This equation represents "line of best fit"

Serves as the "prediction line"

REGRESSION LINE FORMULA

Recall equations of line

Algebra Line: $y = mx + b$

Statistics Line: $y = a + bx$

a = y intercept

b = slope

A = Y INTERCEPT

The value of y if x = 0

**often has no real meaning in context of the data

B = SLOPE

On average, for every increase of one unit in x (explanatory variable, y (response variable) increases or decreases by this amount."

recall slope = $\frac{\Delta y}{\Delta x}$

INTERPRET SLOPES OF REGRESSION LINES

Length = 90 – 3.2 (temp in F)

Income = 22.3 + 0.65(years of service)

Score = 40 +3.24 (attempts)

GOAL OF REGRESSION

- To be able to predict a response based on a value of an explanatory variable

- Example: Predict a student's college GPA based on SAT score

- Example: Predict how much a baby's temperature will go down with x amount of Tylenol

RESIDUAL

Residual: the amount in the y direction that a point is from the regression line

Residual = actual value – predicted value

Positive residual—point above the line

Negative residual—point below the line

LINE OF LEAST SQUARES

The regression line is sometimes called the line of least squares.

The "best fit" minimizes the squares of the residuals of each point from the regression line

EXAMPLE: AGE AND BP

AGE:

43 48 56 61 67 70

Blood Pressure

128 140 135 143 138 152

R

r is the correlation coefficient

*Measures strength of linear relationship

*Can be positive or negative

*Is always a number between -1 and 1

R^2

r^2 is the coefficient of determination

*Measures what % of the variation in y is explained by (or determined by) the variation in x

*Will be given as a %

$1-R^2$

$1 - r^2$ is the coefficient of non-determination

*Measures what % of the variation in y which is explained by chance and other factors

*% of variation in y NOT explained by variation in x

IS LINEAR MODEL APPROPRIATE?

How do we answer this question?

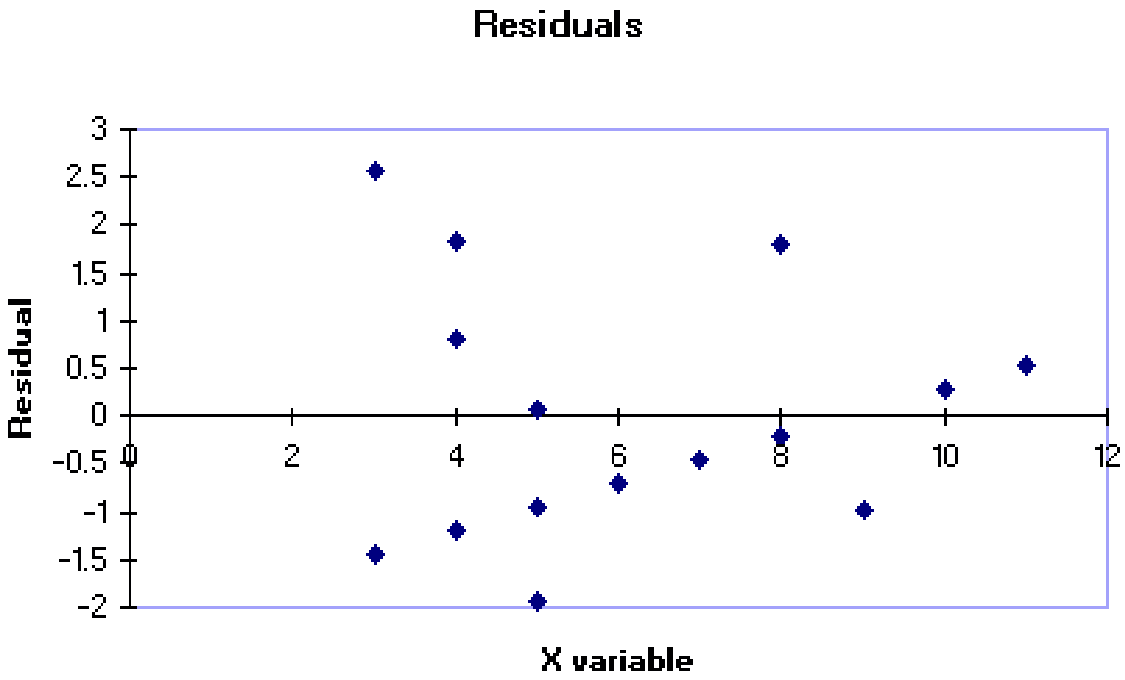
1. Check SCATTERPLOT and r value for strong linearity
2. Look at RESIDUAL PLOT—it should be random

RESIDUAL PLOT

A graph of the residual values compared to the x values.

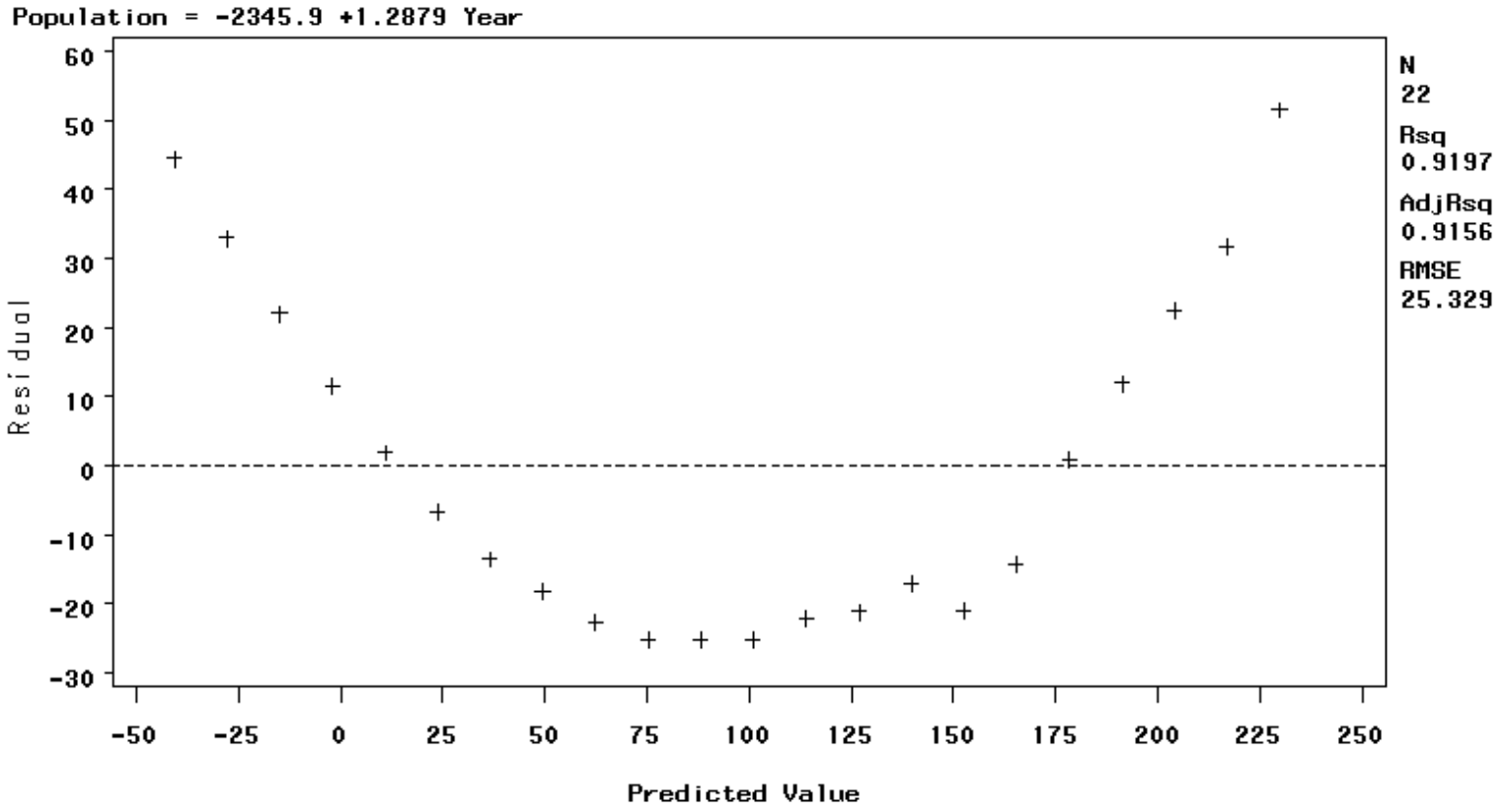
We do not want to see a pattern of residuals increasing or decreasing or fitting any noticeable curve.

GOOD RESIDUAL = RANDOM



BAD RESIDUAL = PATTERN

BAD RESIDUAL = PATTERN



UNUSUAL POINTS?

INFLUENTIAL POINT

A point that strongly affects the regression line. (Usually an outlier, but not always.)

How to check? Remove it from the data and recompute regression line to see if line changes dramatically. Usually have to plug in a point to each regression equation to see if it matters.

OUTLIER

A point in regression analysis can be an outlier in x, an outlier in y or an outlier in both x and y.

How to check? Examine graph to see if one point seems far away from the rest in x or y direction and do outlier test on that variable.

EXTRAPOLATION

The goal of regression is to create a model to predict, but you must be careful how far beyond the range of the original data you predict for.

Predicting y for an x far beyond the range is called extrapolation.

GOOD MODEL = CAUSATION??

If a linear regression model is good (strong r, linear scatterplot, random residuals), that DOES NOT mean that x causes y.

For a researcher to assert that x causes y, the data must have come from a PLANNED, CONTROLLED EXPERIMENT.

R = -.62

What can we say about the level of association between x and y??

*moderate negative linear association

*Only 38% of the variation in y is explained by the variation in x.

RE-EXPRESSING DATA=TRANSFORMATION

We transform data if a linear model is not appropriate.

Transforming data means re-expressing the numbers using algebraic operations that take the "curve" out of the original data.

Examples: take square root, take log or ln,

use reciprocals