Causation, Correlation and Regression

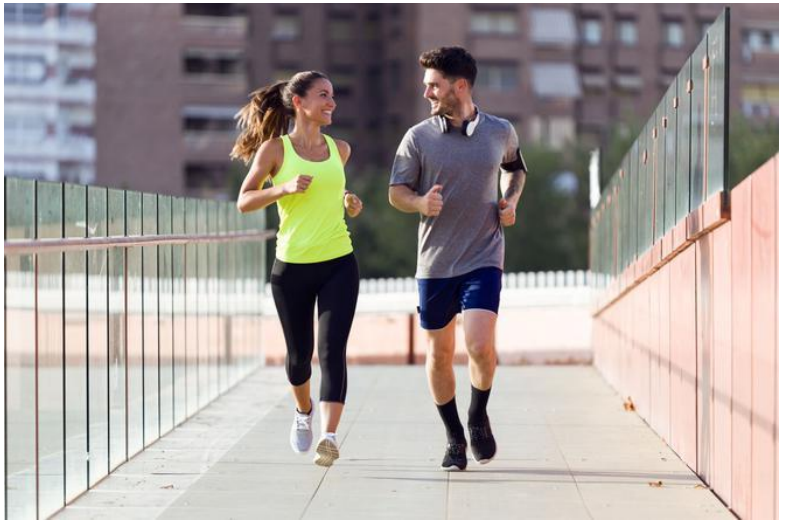
Causation

Causation

Causation is any cause that produces an effect.

This means that when something happens (cause) something else will also always happen(effect).

An example:
When you run you burn calories.



As you can see with the example our cause is running while burning calories is our effect. This is something that is always, because that's how the human body works.

Correlation

Correlation measures the relationship between two things.

Positive correlations happen when one thing goes up, and another thing goes up as well.
An example: When the demand for a product is high, the price may go up. As you can see, because the demand is high the price may be high.

Negative correlations occur when the opposite happens. When one thing goes up, and another goes down.

A correlation tells us that two variables are related, but we cannot say anything about whether one caused the other.

Correlation

Correlations happen when:
A causes B

B causes A

A and B are consequences of a common cause, but do not cause each other

There is no connection between A and B, the correlation is coincidental

Causation and Correlation

Causation and correlation can happen at the same time. But having a correlation does not always mean you have a causation.

A good example of this:
There is a positive correlation between the number of firemen fighting a fire and the size of the fire. This means the more people at the fire, tends to reflect how big the fire is. However, this doesn't mean that bringing more firemen will cause the size of the fire to increase.

Correlation or Causation?

As people's happiness level increases, so does their helpfulness.

This would be a correlation.
Just because someone is happy does not always mean that they will become more helpful. This just usually tends to be the case.

Correlation or Causation?

Dogs pant to cool themselves down.

This would be a causation.

When a dog needs to cool itself down it will pant. This is not something that tends to happen, it is something that is always true.

Correlation or Causation?

Among babies, those who are held more tend to cry less.

This would be a correlation.
Just because a baby is held often does not mean that it will cry less. This just usually tends to be the case.

Let's think of our own

Correlation:

Causation:

Quick Review

Causation is any cause that produces an effect.

Correlation measure the relationship between two things.

Correlation

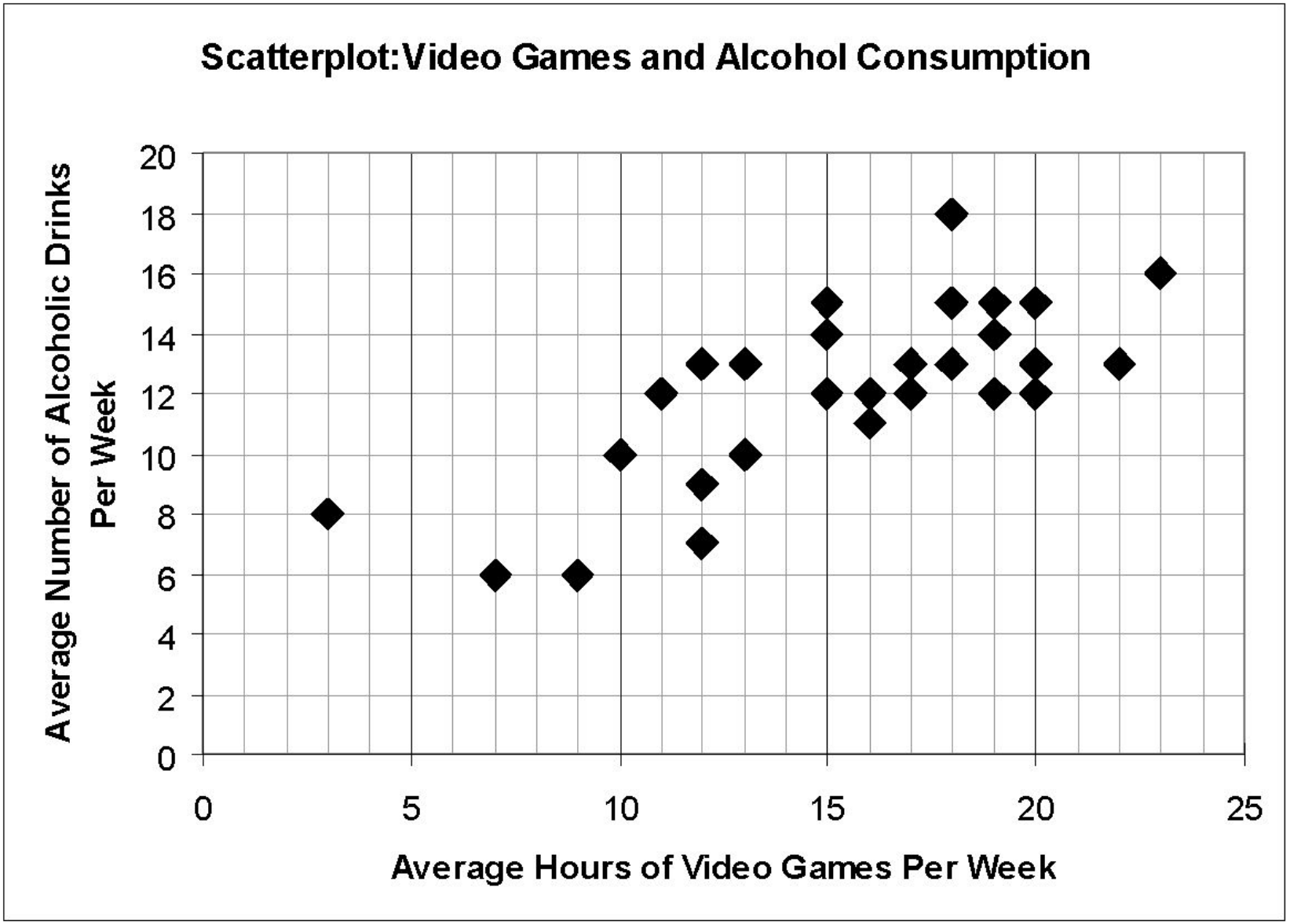
The Question

- Are two variables related?
  - Does one increase as the other increases?
    - e. g. skills and income
  - Does one decrease as the other increases?
    - e. g. health problems and nutrition
- How can we get a numerical measure of the degree of relationship?

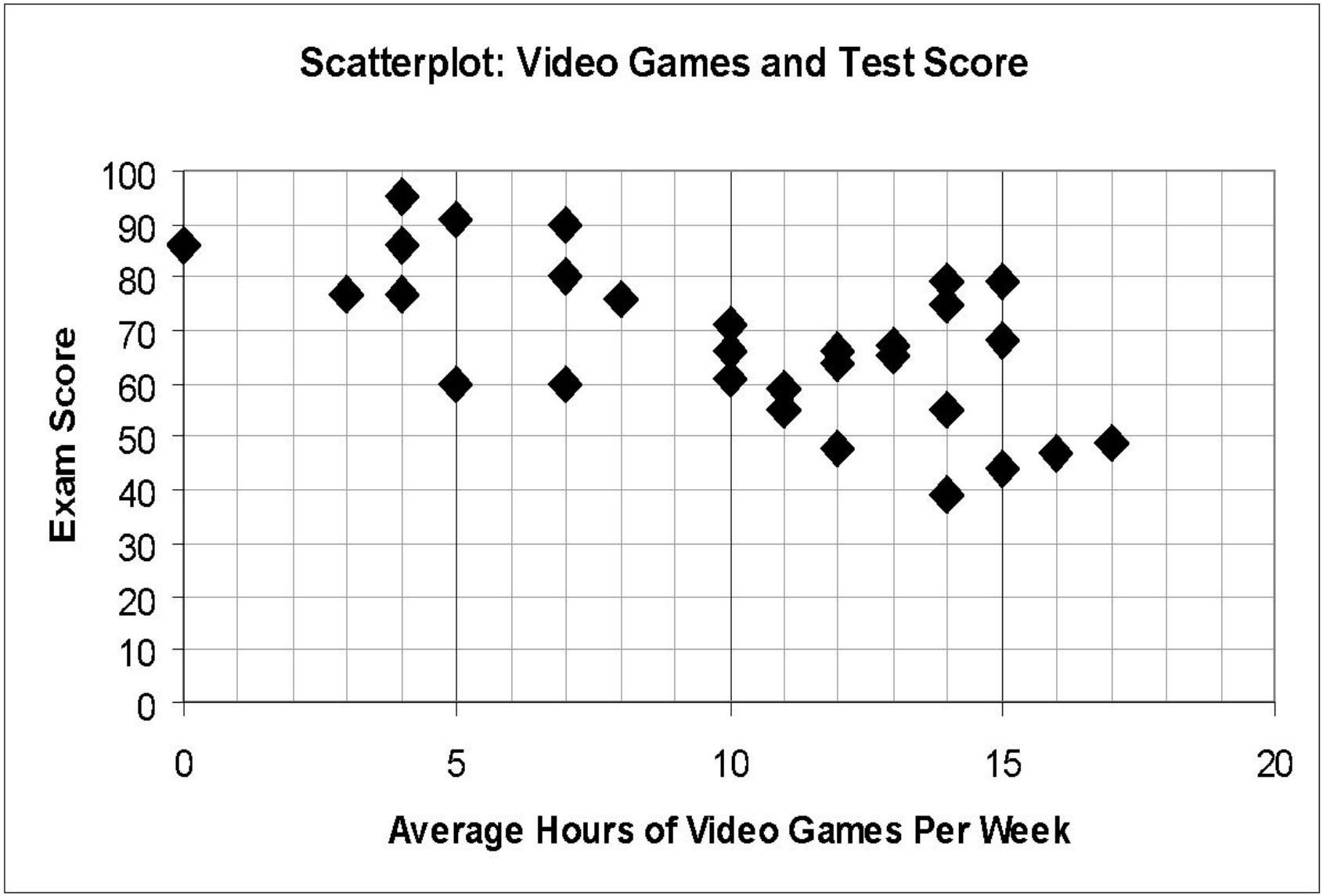
Scatterplots

- Graphically depicts the relationship between two variables in two dimensional space.

Direct Relationship



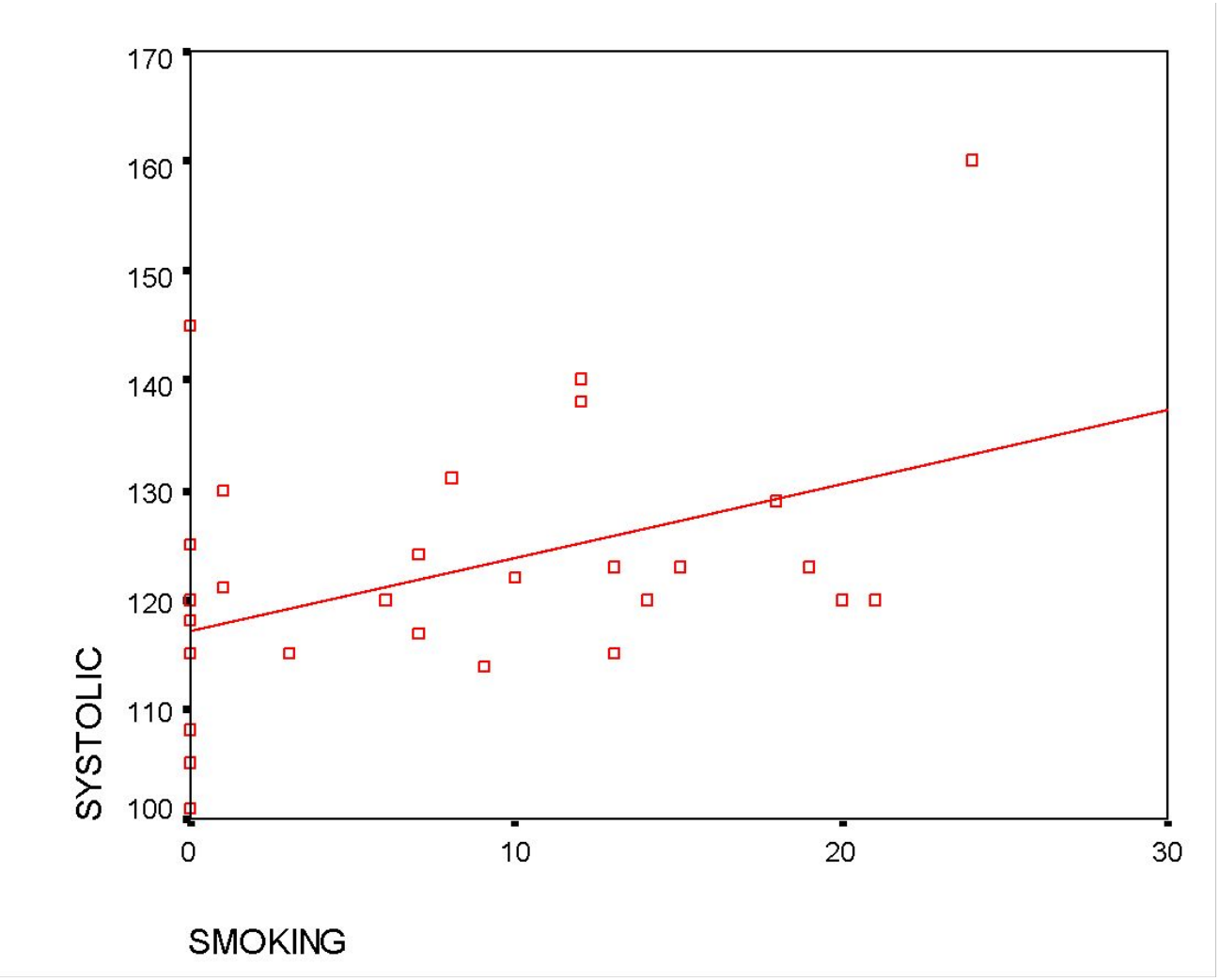
Inverse Relationship



An Example

- Does smoking cigarettes increase systolic blood pressure?
- Plotting number of cigarettes smoked per day against systolic blood pressure
  - Fairly moderate relationship
  - Relationship is positive

Trend?



Smoking and BP

- Note relationship is moderate, but real.
- Why do we care about relationship?
  - What would conclude if there were no relationship?
  - What if the relationship were near perfect?
  - What if the relationship were negative?

Heart Disease and Cigarettes

- Data on heart disease and cigarette smoking in 21 developed countries Data have been rounded for computational convenience.
  - The results were not affected.

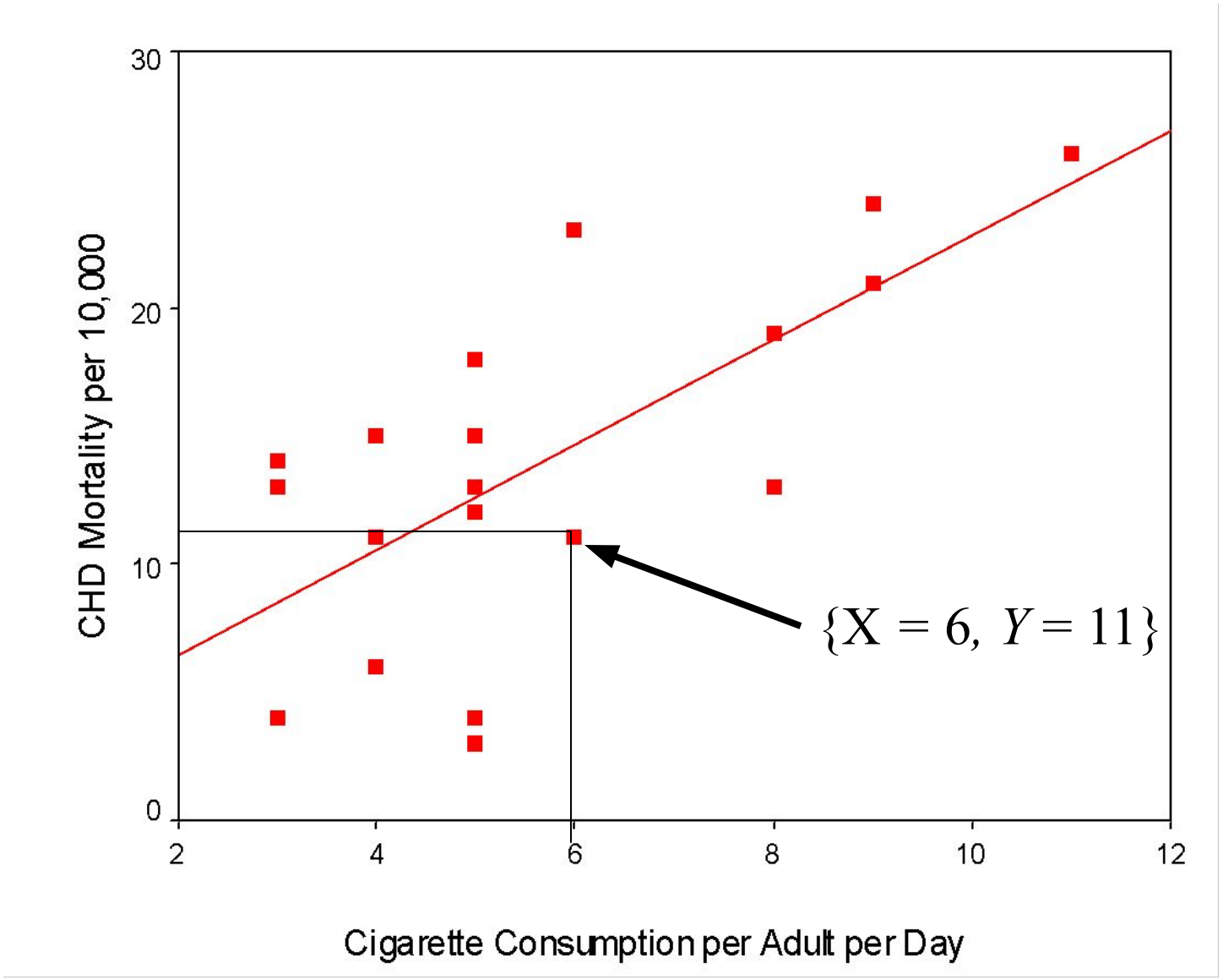
The Data

Surprisingly, the U.S. is the first country on the list--the country with the highest consumption and highest mortality.

Country	Cigarettes	CHD
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14

Scatterplot of Heart Disease

- CHD Mortality goes on Y axis
  - Why?
- Cigarette consumption on X axis
  - Why?
- What does each dot represent?
- Best fitting line included for clarity


30
20
10
0
CHD Mortality per 10,000
2
4
6
8
10
12
Cigarette Consumption per Adult per Day
{X = 6, Y = 11}

What Does the Scatterplot Show?

- As smoking increases, so does coronary heart disease mortality.
- Relationship looks strong
- Not all data points on line.
  - This gives us “residuals” or “errors of prediction”
    - To be discussed later

Correlation

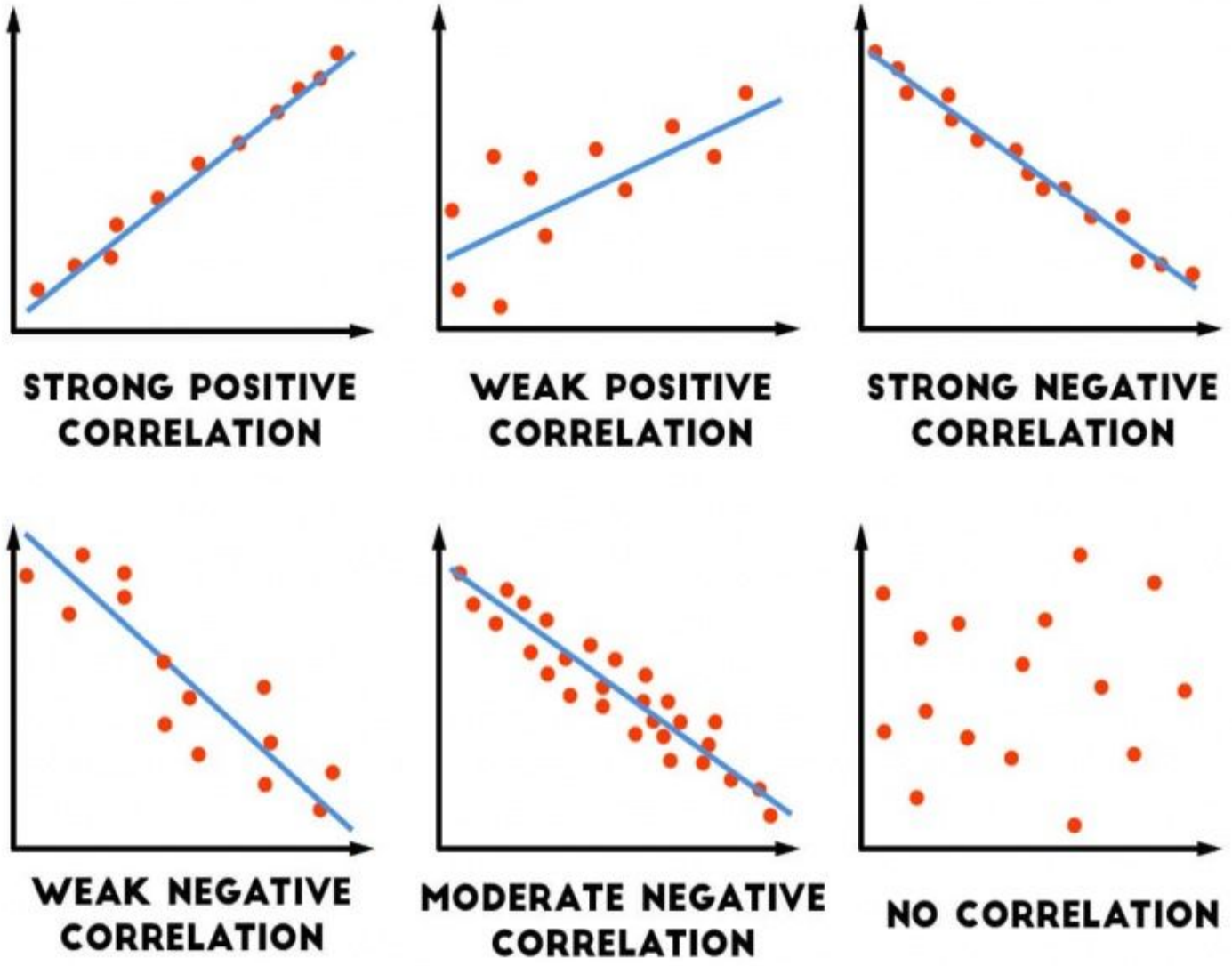
- Co-relation
- The relationship between two variables
- Measured with a correlation coefficient
- Most popularly seen correlation coefficient: Pearson Product-Moment Correlation

Types of Correlation

- Positive correlation
  - High values of X tend to be associated with high values of Y.
  - As X increases, Y increases
- Negative correlation
  - High values of X tend to be associated with low values of Y.
  - As X increases, Y decreases
- No correlation
- No consistent tendency for values on Y to increase or decrease as X increases

Correlation Coefficient

- A measure of degree of relationship.
- Between 1 and -1
- Sign refers to direction.
- Based on covariance
  - Measure of degree to which large scores on X go with large scores on Y, and small scores on X go with small scores on Y


STRONG POSITIVE CORRELATION
WEAK POSITIVE CORRELATION
STRONG NEGATIVE CORRELATION
WEAK NEGATIVE CORRELATION
MODERATE NEGATIVE CORRELATION
NO CORRELATION

Covariance

- The formula for co-variance is:

$$Cov_{XY} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N - 1}$$

- How this works, and why?
- When would cov_{XY} be large and positive? Large and negative?

Example

Country	X (Cig.)	Y (CHD)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})*(Y-\bar{Y})$
1	11	26	5.05	11.48	57.97
2	9	21	3.05	6.48	19.76
3	9	24	3.05	9.48	28.91
4	9	21	3.05	6.48	19.76
5	8	19	2.05	4.48	9.18
6	8	13	2.05	-1.52	-3.12
7	8	19	2.05	4.48	9.18
8	6	11	0.05	-3.52	-0.18
9	6	23	0.05	8.48	0.42
10	5	15	-0.95	0.48	-0.46
11	5	13	-0.95	-1.52	1.44
12	5	4	-0.95	-10.52	9.99
13	5	18	-0.95	3.48	-3.31
14	5	12	-0.95	-2.52	2.39
15	5	3	-0.95	-11.52	10.94
16	4	11	-1.95	-3.52	6.86
17	4	15	-1.95	0.48	-0.94
18	4	6	-1.95	-8.52	16.61
19	3	13	-2.95	-1.52	4.48
20	3	4	-2.95	-10.52	31.03
21	3	14	-2.95	-0.52	1.53
Mean	5.95	14.52			
SD	2.33	6.69			
Sum					222.44

Example

$$Cov_{cig.\&CHD} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N - 1} = \frac{222.44}{21 - 1} = 11.12$$

- What the heck is a covariance?
- I thought we were talking about correlation?

Correlation Coefficient

- Pearson's Product Moment Correlation
- Symbolized by *r*
- Covariance ÷ (product of the 2 SDs)

$$r = \frac{Cov_{XY}}{s_X s_Y}$$

- Correlation is a standardized covariance

Calculation for Example

- $Cov_{XY} = 11.12$
- $s_X = 2.33$
- $s_Y = 6.69$

$$r = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{11.12}{(2.33)(6.69)} = \frac{11.12}{15.59} = .713$$

Example

- Correlation = .713
- Sign is positive
  - Why?
- If sign were negative
  - What would it mean?
  - Would not change the *degree* of relationship.

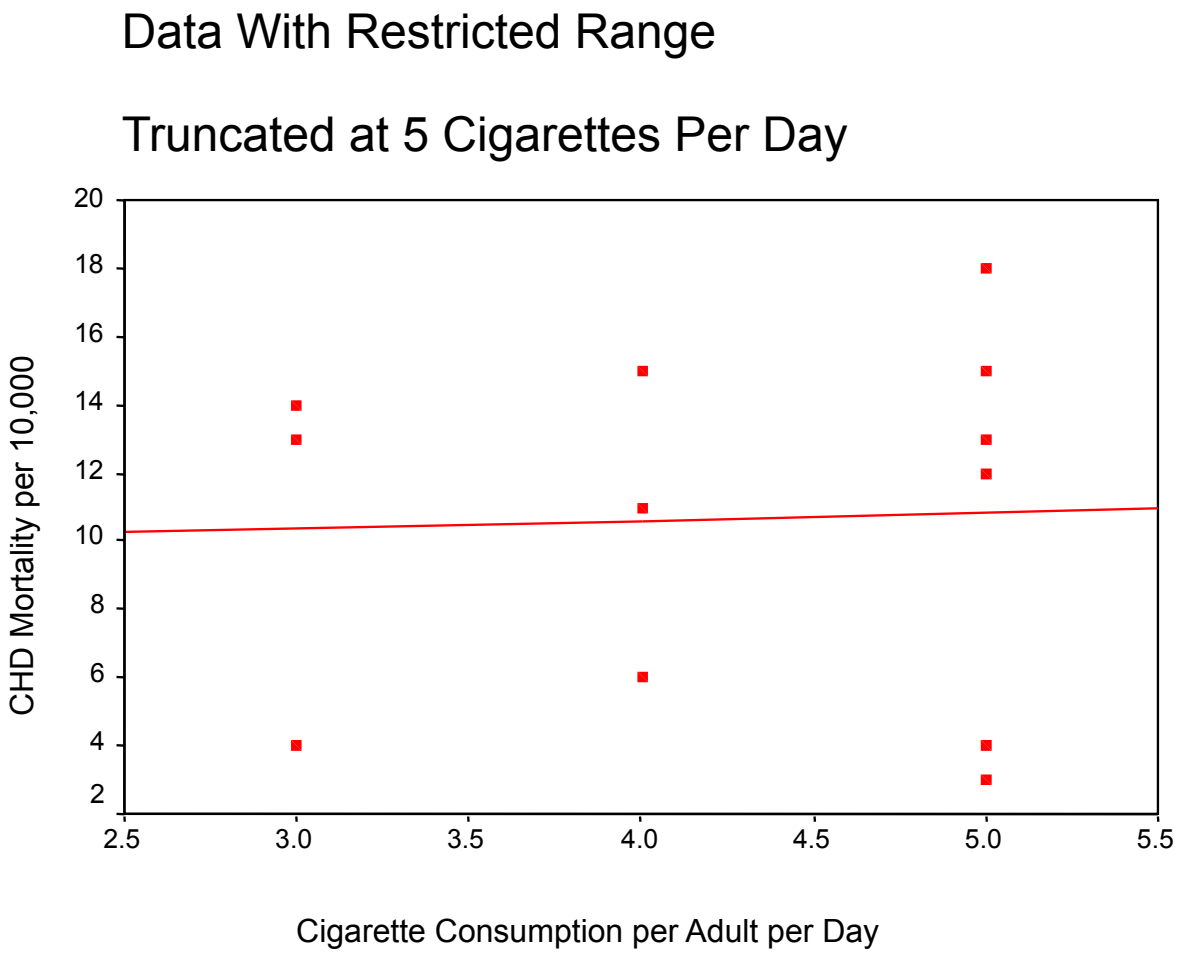
Factors Affecting *r*

- Range restrictions
  - Looking at only a small portion of the total scatter plot (looking at a smaller portion of the scores' variability) **decreases** *r*.
  - Reducing variability reduces *r*
- Nonlinearity
  - The Pearson r measures the degree of **linear** relationship between two variables
  - If a strong non-linear relationship exists, r will provide a low, or at least inaccurate measure of the true relationship.

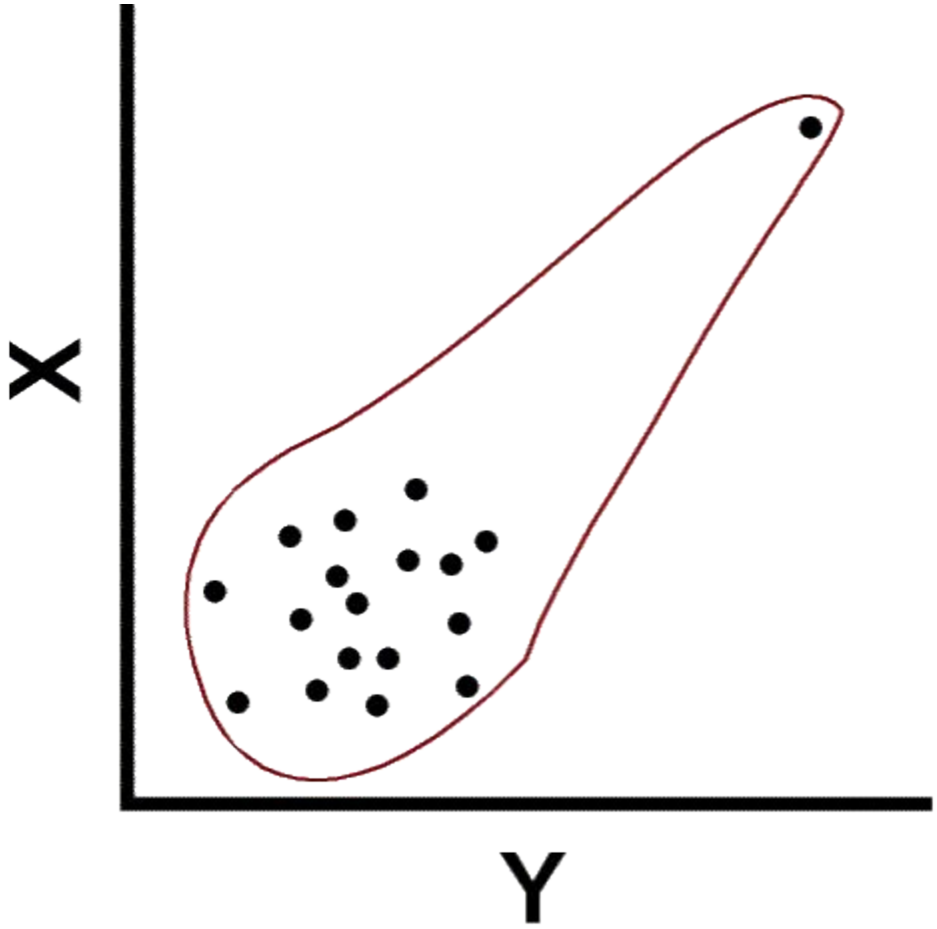
Factors Affecting *r*

- Outliers
  - Overestimate Correlation
  - Underestimate Correlation

Countries With Low Consumptions



Outliers



Testing Correlations

- So you have a correlation. Now what?
- In terms of magnitude, how big is big?
  - Small correlations in large samples are "big."
  - Large correlations in small samples aren't always "big."
- Depends upon the magnitude of the correlation coefficient

AND

- The size of your sample.

Regression

- „Regression” refers to the process of fitting a simple line to datapoints, Historically, linear regression was first used to explain the height of men by the height of their fathers.

What is regression?

- How do we predict one variable from another?
- How does one variable change as the other changes?
- Influence

Linear Regression

- A technique we use to predict the most likely score on one variable from those on another variable
- Uses the *nature of the relationship* (i.e. correlation) between two variables to *enhance* your prediction

Linear Regression: Parts

- Y - the variables you are predicting
  - i.e. dependent variable
- X - the variables you are using to predict
  - i.e. independent variable
- $\hat{Y}$ - your predictions (also known as Y')

Why Do We Care?

- We may want to make a prediction.
- More likely, we want to understand the relationship.
  - How fast does CHD mortality rise with a one unit increase in smoking?
  - Note: we speak about predicting, but often don't actually predict.

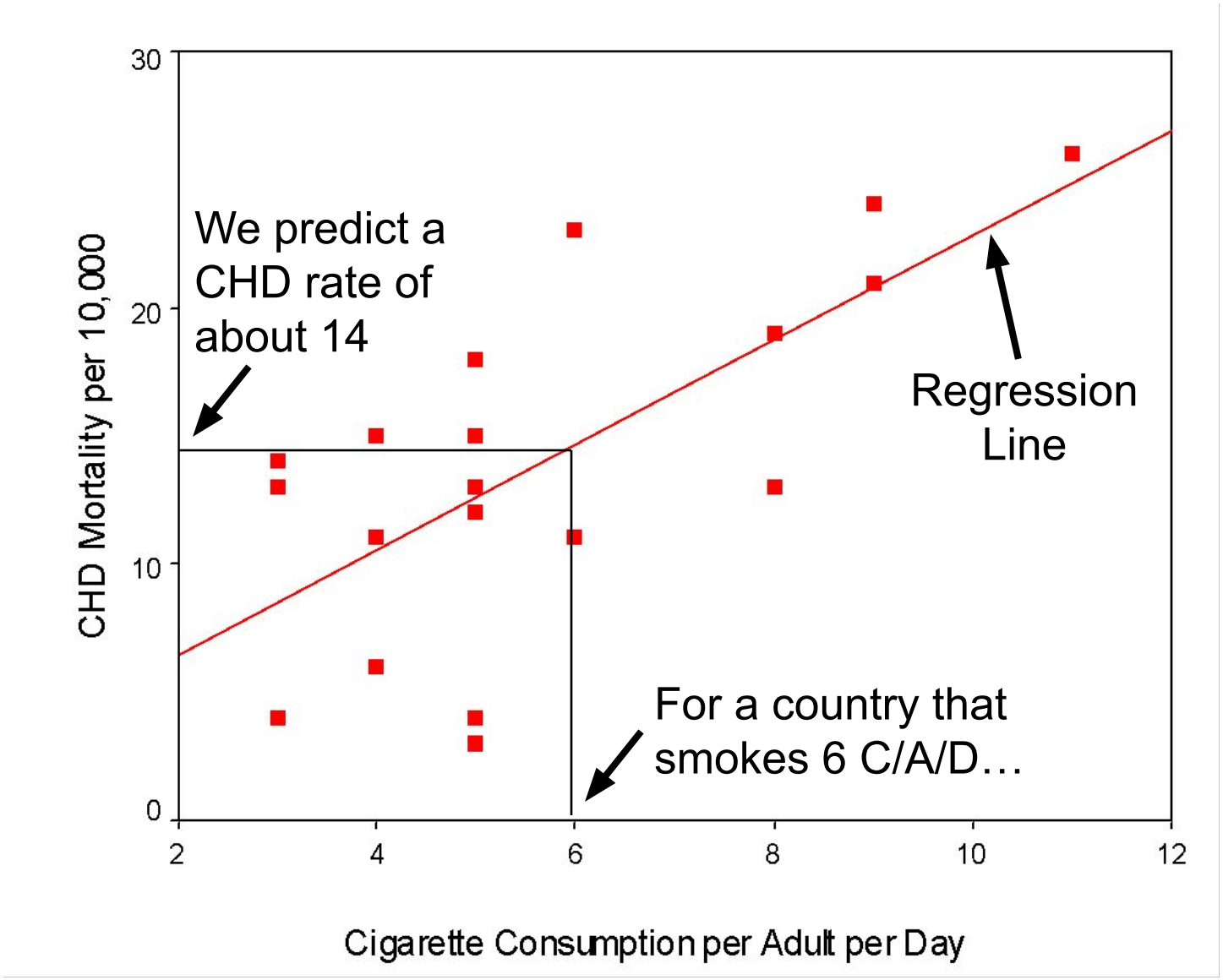
An Example

- Cigarettes and CHD Mortality again
- Data repeated on next slide
- We want to predict level of CHD mortality in a country averaging 10 cigarettes per day.

The Data

Based on the data we have what would we predict the rate of CHD be in a country that smoked 10 cigarettes on average?
First, we need to establish a prediction of CHD from smoking…

Country	Cigarettes	CHD
1	11	26
2	9	21
3	9	24
4	9	21
5	8	19
6	8	13
7	8	19
8	6	11
9	6	23
10	5	15
11	5	13
12	5	4
13	5	18
14	5	12
15	5	3
16	4	11
17	4	15
18	4	6
19	3	13
20	3	4
21	3	14


30
20
10
0
CHD Mortality per 10,000
We predict a CHD rate of about 14
Regression Line
For a country that smokes 6 C/A/D…
2
4
6
8
10
12
Cigarette Consumption per Adult per Day

Regression Line

- Formula

$$\hat{Y} = bX + a$$

  - $\hat{Y}$ = the predicted value of *Y* (e.g. CHD mortality)
  - X = the predictor variable (e.g. average cig./adult/country)

Regression Coefficients

- "Coefficients" are *a* and *b*
- *b* = slope
  - Change in predicted *Y* for one unit change in *X*
- *a* = intercept
  - value of $\hat{Y}$ when $X = 0$

Calculation

- Slope

$$b = \frac{\text{cov}_{XY}}{s_X^2} \text{ or } b = r\left[\frac{s_y}{s_x}\right]$$

$$\text{or } b = \frac{N\sum XY - \sum X \sum Y}{\left[N\sum X^2 - (\sum X)^2\right]}$$

- Intercept

$$a = \overline{Y} - b\overline{X}$$

For Our Data

- $Cov_{XY} = 11.12$
- $s^2_X = 2.33^2 = 5.447$
- $b = 11.12/5.447 = 2.042$
- $a = 14.524 - 2.042*5.952 = 2.32$

Note:

- The values we obtained are shown on printout.
- The intercept is the value in the *B* column labeled “constant”
- The slope is the value in the *B* column labeled by name of predictor variable.

Making a Prediction

- Second, once we know the relationship we can predict

$$\hat{Y} = bX + a = 2.042X + 2.367$$

$$\hat{Y} = 2.042 * 10 + 2.367 = 22.787$$

- We predict 22.77 people/10,000 in a country with an average of 10 C/A/D will die of CHD

Accuracy of Prediction

- Finnish smokers smoke 6 C/A/D
- We predict:

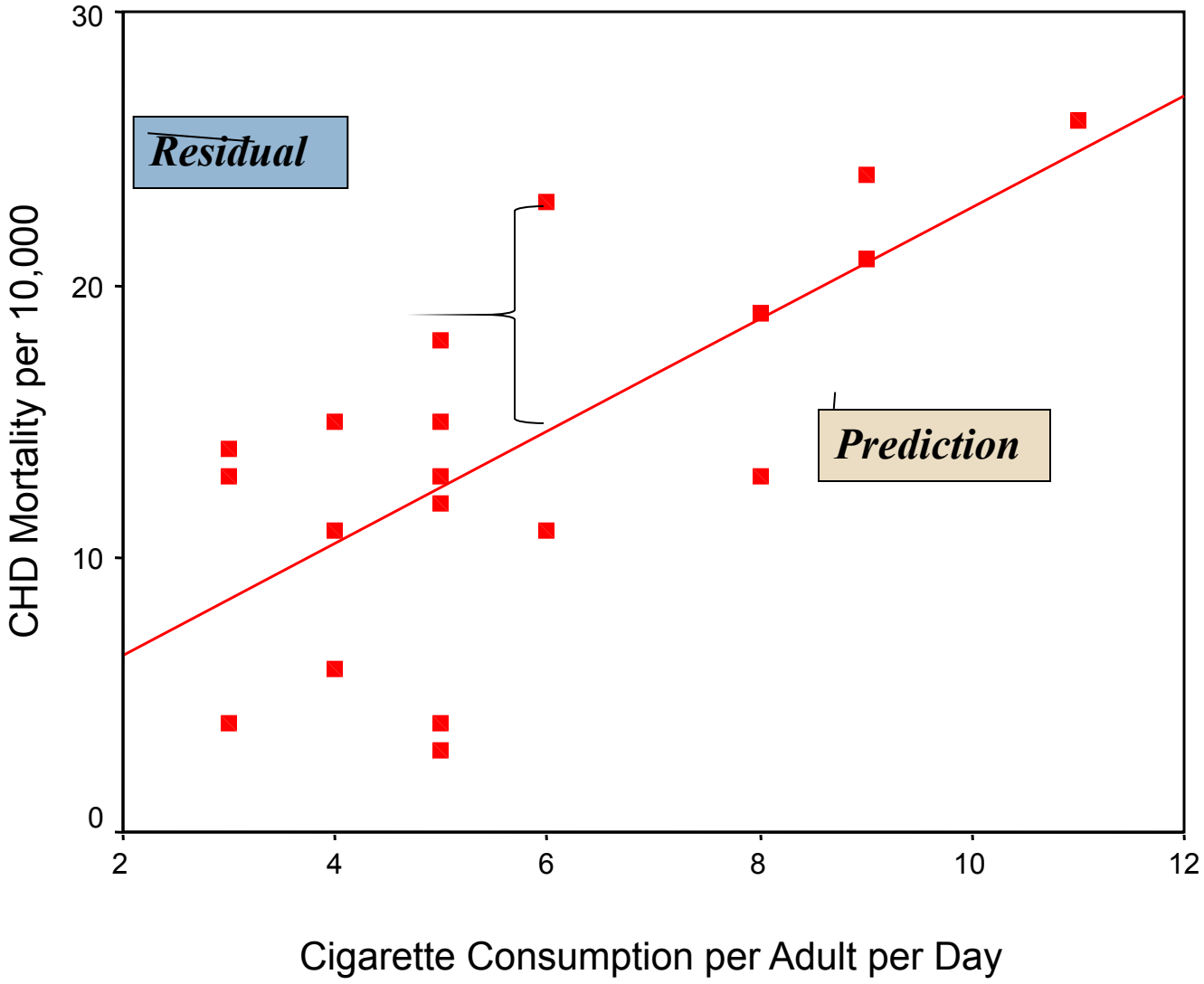
$$\hat{Y} = bX + a = 2.042X + 2.367$$

$$\hat{Y} = 2.042 * 6 + 2.367 = 14.619$$

- They actually have 23 deaths/10,000
- Our error ("residual") =

23 - 14.619 = 8.38

  - a large error


Residual
Prediction
CHD Mortality per 10,000
30
20
10
0
2
4
6
8
10
12
Cigarette Consumption per Adult per Day

Residuals

- When we predict $\hat{Y}$ for a given X, we will sometimes be in error.
- $Y - \hat{Y}$ for any X is a an ***error of estimate***
- Also known as: a ***residual***
- We want to $\Sigma(Y - \hat{Y})$ as small as possible.
- BUT, there are infinitely many lines that can do this.
- Just draw ANY line that goes through the mean of the X and Y values.
- Minimize Errors of Estimate… How?

Minimizing Residuals

- Again, the problem lies with this definition of the mean:

$$\sum (X - \bar{X}) = 0$$

- So, how do we get rid of the 0's?
- Square them.

Regression Line: A Mathematical Definition

- The regression line is the line which when drawn through your data set produces the smallest value of:

$$\sum (Y - \hat{Y})^2$$

- Called the Sum of Squared Residual or $SS_{residual}$
- Regression line is also called a "least squares line."