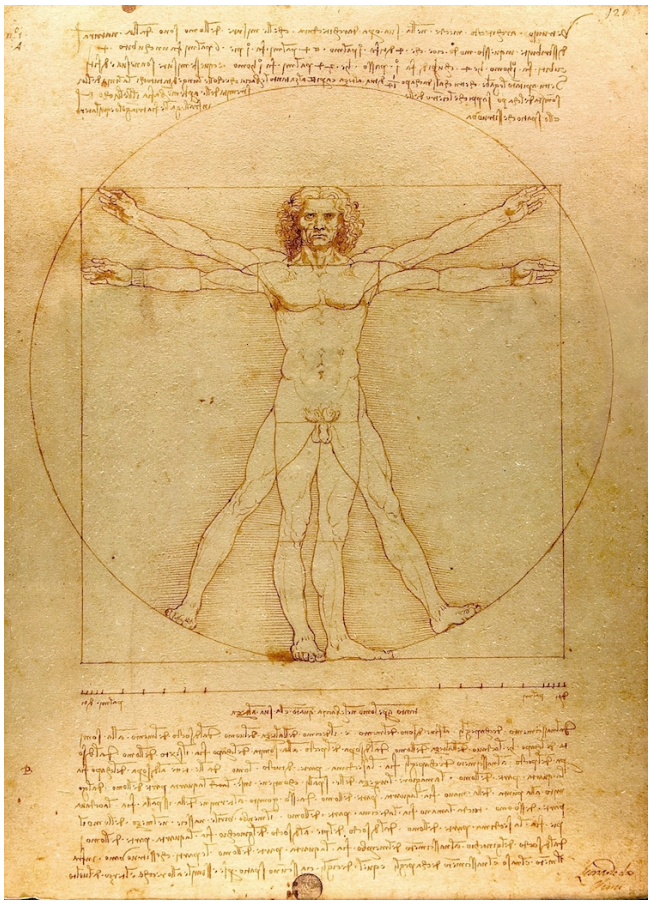
Paired Data

- Bivariate Data that are coupled or matched together. They are not independent.

Example:

- Leonardo da Vinci's *Vitruvian Man*.



- The outstretched arms and legs within circles and square.
- Ideal human proportions described by ancient Roman architect Vitrivius: height is same as length of arm span.

Paired Data

Key Tools to understand paired data:

- Plot to gauge relationship.
- Correlation between the variables.
- Trends

Consider `fat` dataset in `UsingR` package. The dataset contains body dimensions of 250 males.

```
> require(UsingR)
> names(fat)
 [1] "case"          "body.fat"      "body.fat.siri"
 [4] "density"       "age"           "weight"
 [7] "height"        "BMI"           "ffweight"
[10] "neck"          "chest"         "abdomen"
[13] "hip"           "thigh"         "knee"
[16] "ankle"         "bicep"         "forearm"
[19] "wrist"
```

- Suppose we are interested in relation between neck and wrist.

Paired Data

We can first compare averages in two ways:

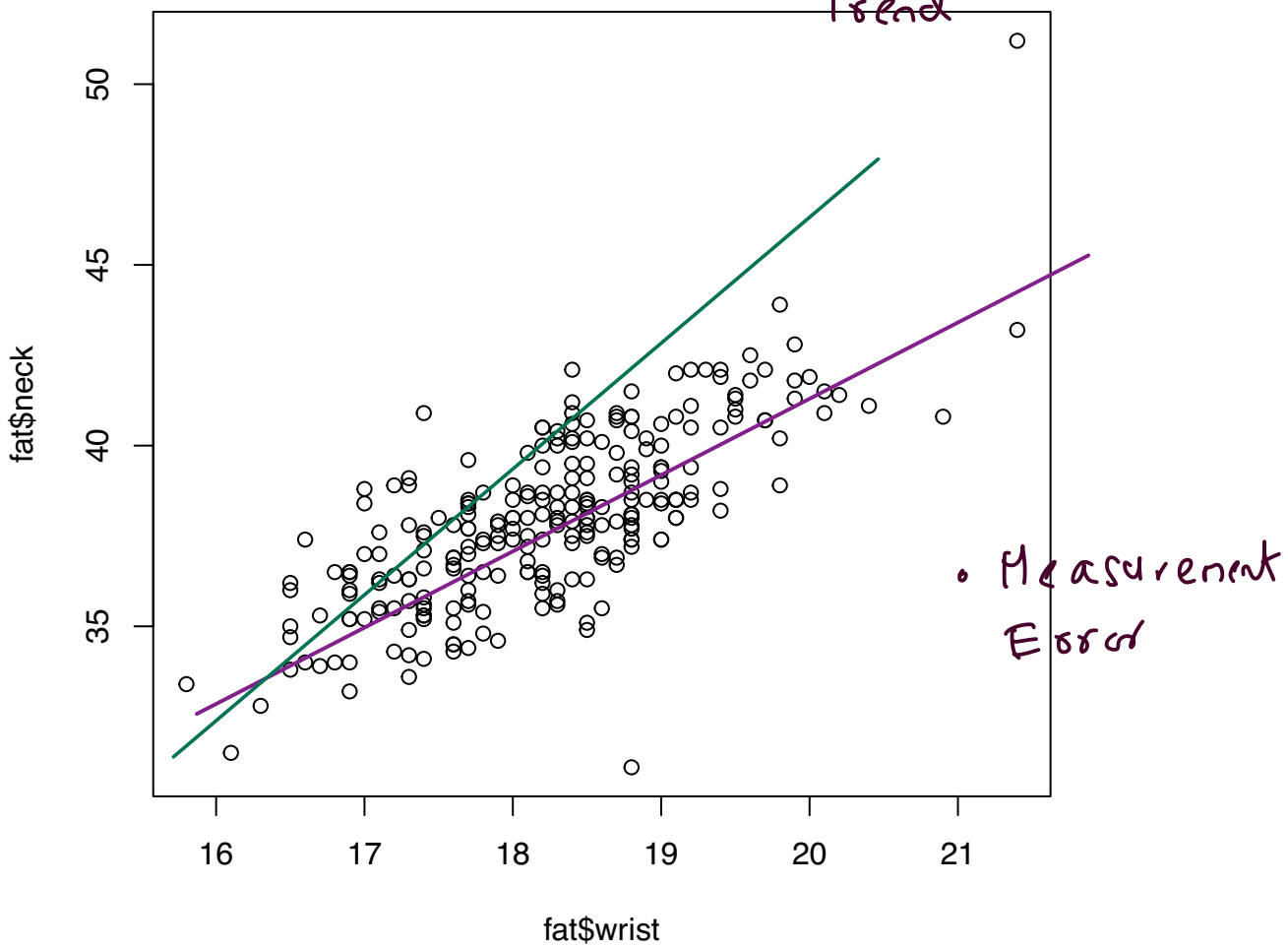
```
> z = mean(fat$neck)/mean(fat$wrist)
> z
[1] 2.084068
> y = mean(fat$neck/fat$wrist)
> y
[1] 2.084477
```

Is there another measure?

We can plot to check:

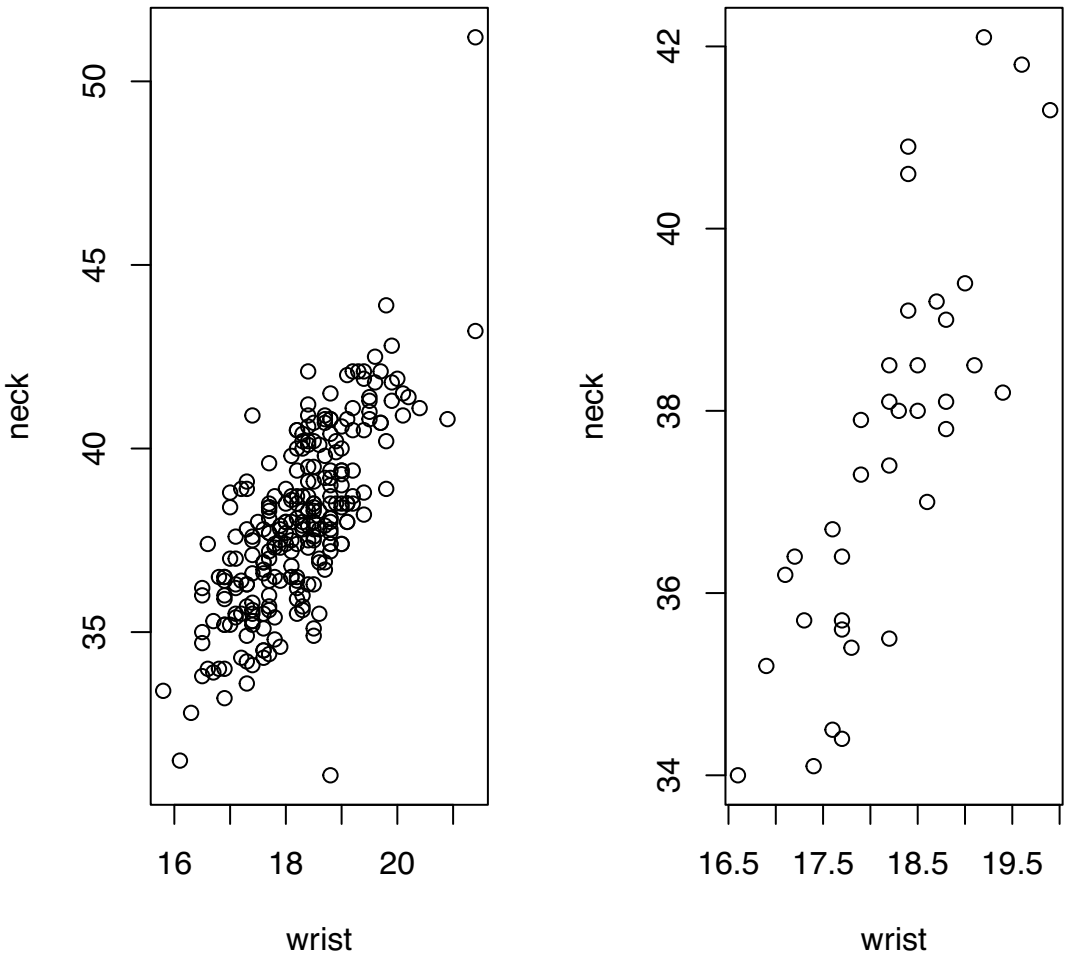
```
> plot(fat$wrist, fat$neck)
```

Understand
- Appropriate Trend
- Measurement Error



Paired Data: `fat` dataset in UsingR

```
> par(mfrow=c(1,2))
> plot(neck~wrist, data=fat)
> plot(neck~wrist, data=fat, subset=20<=age &age <30)
```



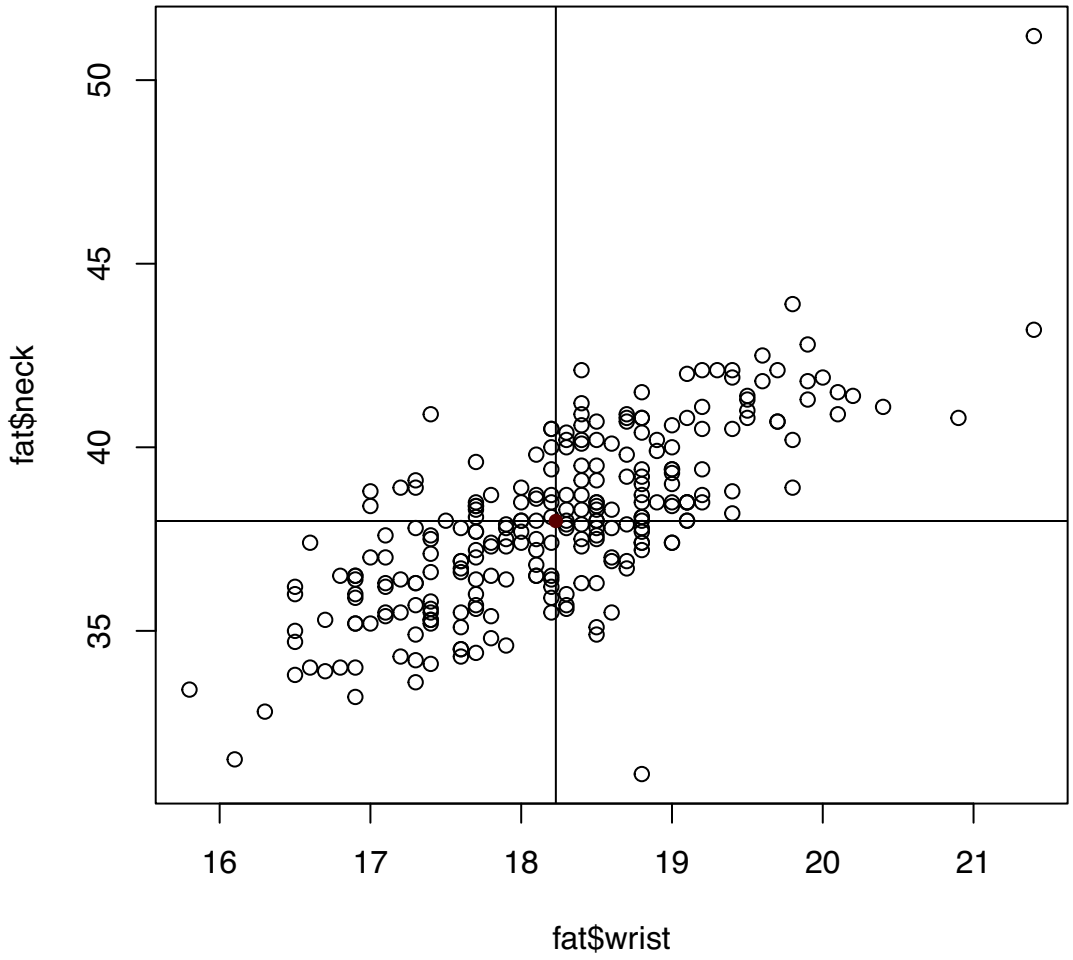
The variables seem related and also by a linear relationship

Suppose x = fat$wrist, y = fat$neck : $y = \beta_1 x + \beta_0$

what is β_1?

Data in four regions by means

```
> plot(fat$wrist, fat$neck)
> abline(v=mean(fat$wrist))
> abline(h=mean(fat$neck))
> points(mean(fat$wrist), mean(fat$neck),
+ pch=16, col=rgb(.35,0,0))
```



- Understand data by those above average values and those below.
- If related then most of data should be in first and third box.

Pearson Correlation Coefficient

Covariance measures the difference between the two variables in the four regions. Suppose we have a dataset $\{(x_i, y_i) : 1 \leq i \leq n\}$ then

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Data with strong linear relationship $(x_i - \bar{x})(y_i - \bar{y})$ will have the same sign. (i.e if data lies in first and third box or in second and fourth box).
- In such cases covariance will be large in absolute value.

let $S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

Correlation is Covariance in standardised scale. Suppose we have a dataset $\{(x_i, y_i) : 1 \leq i \leq n\}$ then

$$\text{Cor}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{S_x} \right) \left(\frac{(y_i - \bar{y})}{S_y} \right)$$

- $\text{Cor}(x, y)$ is between -1 and 1.
- $\text{Cor}(x, y) \in \{1, -1\}$ indicates perfect linear relationship.
- $\text{Cor}(x, y) = 0$ indicates no linear relationship.

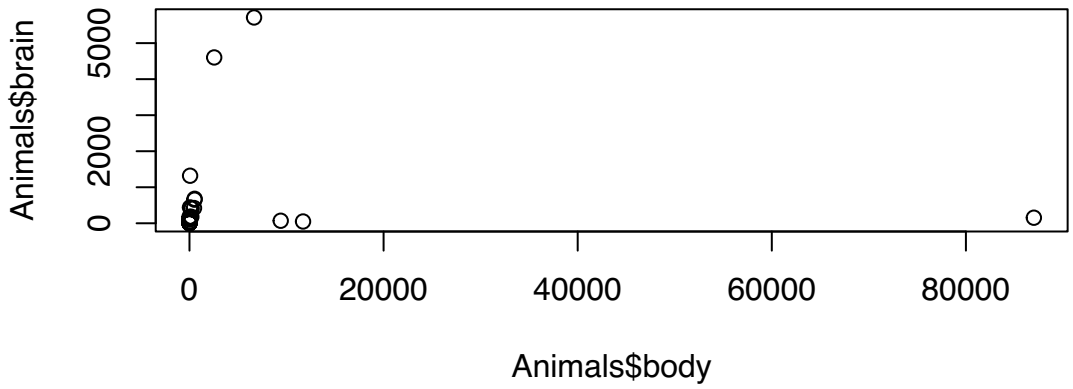
Correlation is a measure of how close the linear relationship is.

Paired Data: `fat` dataset in UsingR

```
> cor(fat$wrist, fat$neck)
[1] 0.7448264
```

Pearson Correlation Coefficient

```
> require(MASS)
> plot(Animals$body,Animals$brain)
```




```
> cor(Animals$body,Animals$brain)
[1] -0.005341163
```

One way is to exclude the outliers. Another method is to transform the data set by placing data in order and assigning a rank. Use `rank`.

```
> cor(rank(Animals$body), rank(Animals$brain))
[1] 0.7162994
```

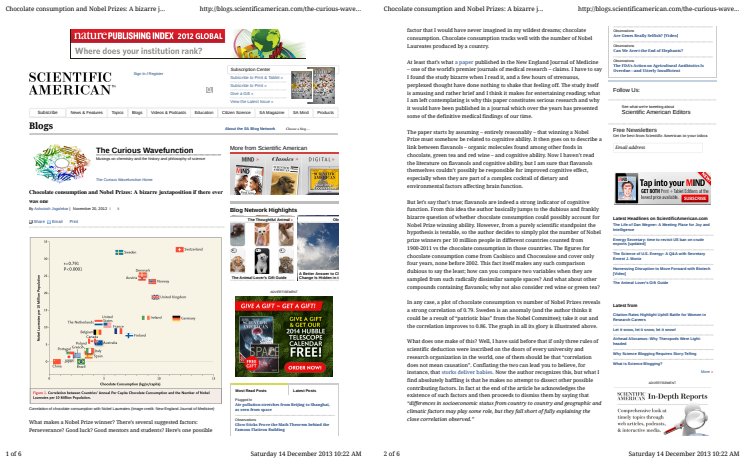
Suppose we have a dataset $\{(x_i, y_i) : 1 \leq i \leq n\}$ then first rank them to get .$\{(r_{x_i}, r_{y_i}) : 1 \leq i \leq n\}$

$$\text{Spearman Correlation}(x, y) = \text{Cor}(r_x, r_y)$$

This is useful in measurement of relationship of monotonic data and not restricted to linear.

```
> cor(Animals$body, Animals$brain, method="spearman")
[1] 0.7162994
```

Chocolates and Noble Prizes



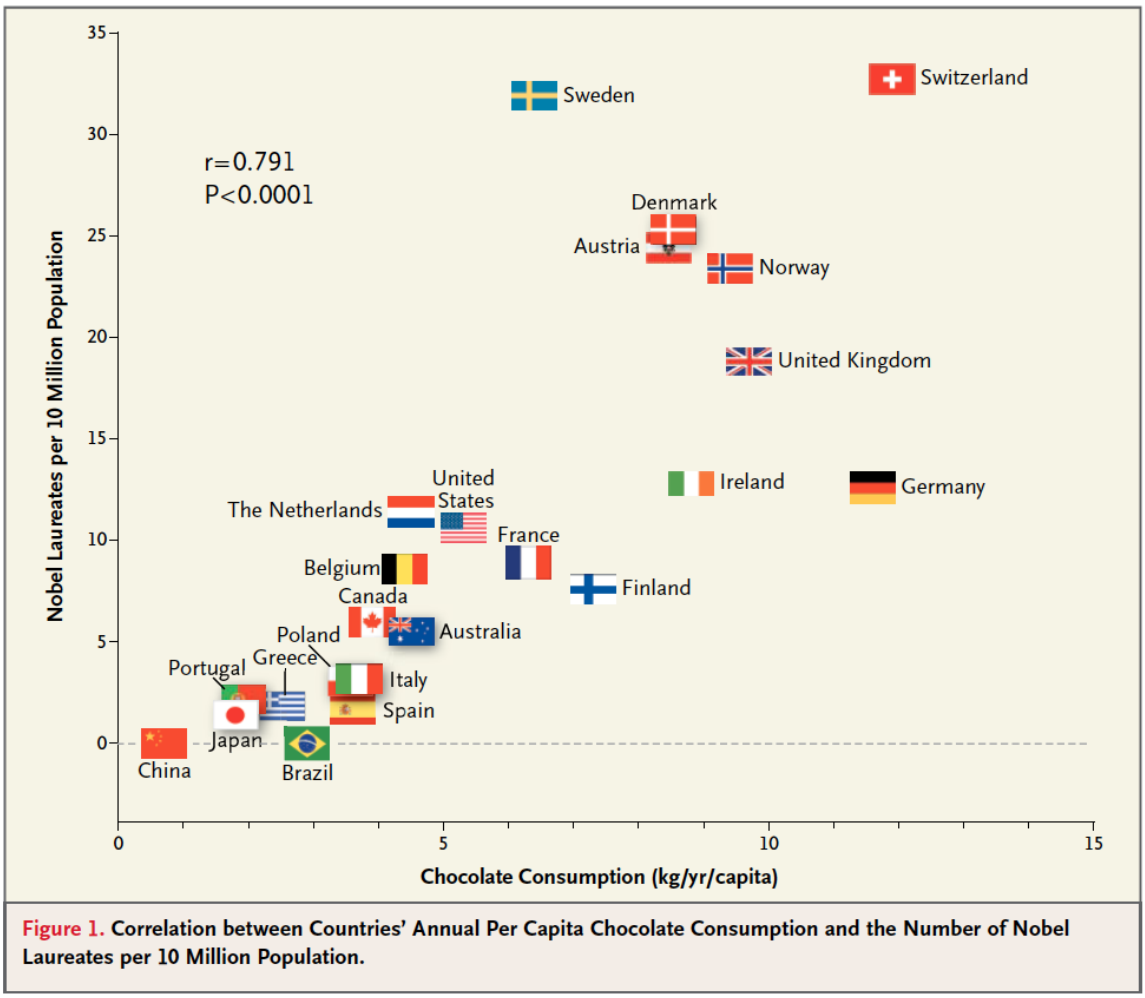
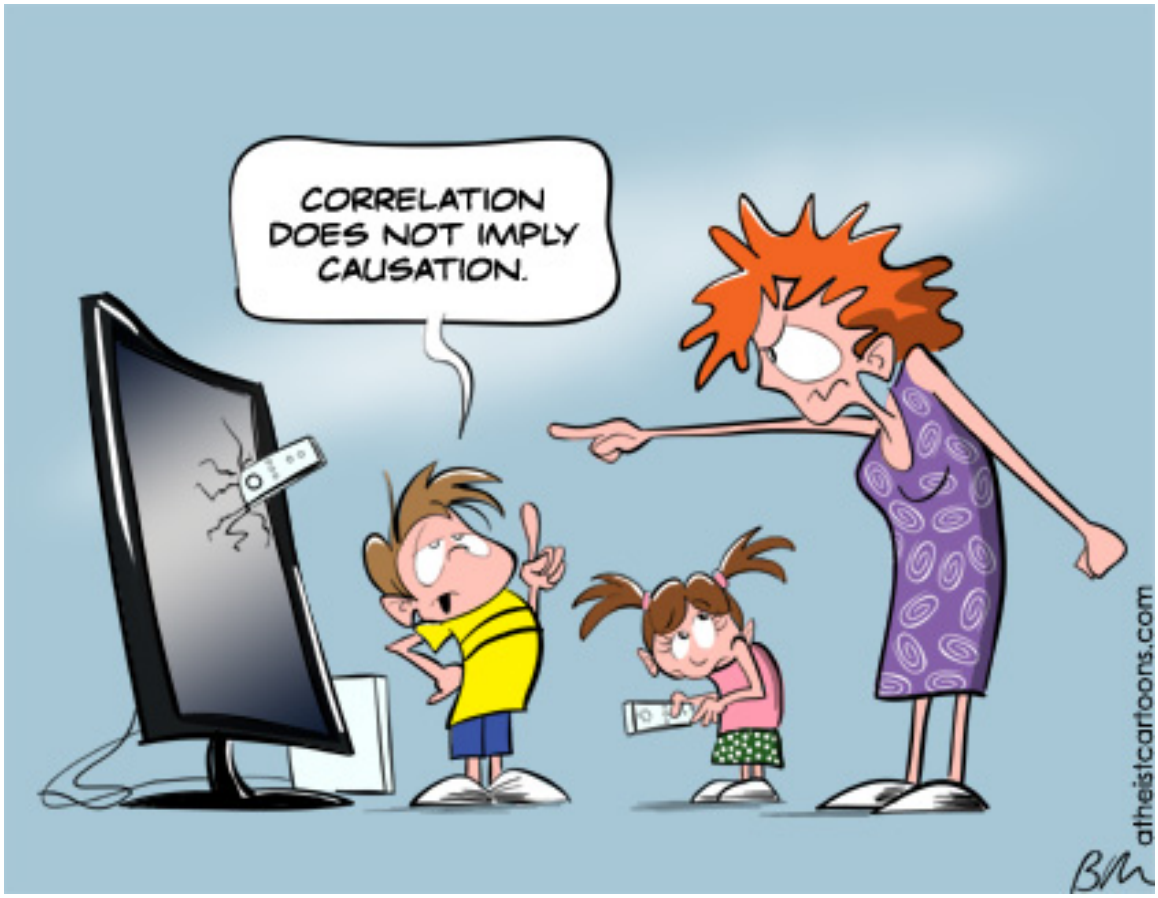


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Noticed: Countries with more per capita chocolate consumption have more per capita Nobel laureates.

Conclude: Chocolate consumption cause better scientific research !

Correlation



- Spurious: Facebook Users and Marks of users
- Causality: Smoking and lung cancer, Wine and heart risk.
- Pearson correlation coefficient is a measure of the linearity of the (possible) relationship between two variables X and Y.
- Even if correlation coefficient is high, it does not mean there is causal relationship between X and Y. Does not tell you cause and effect ?
- Care to be taken when used for predictive purposes.
- Causality: Domain Knowledge, design a good control experiment.