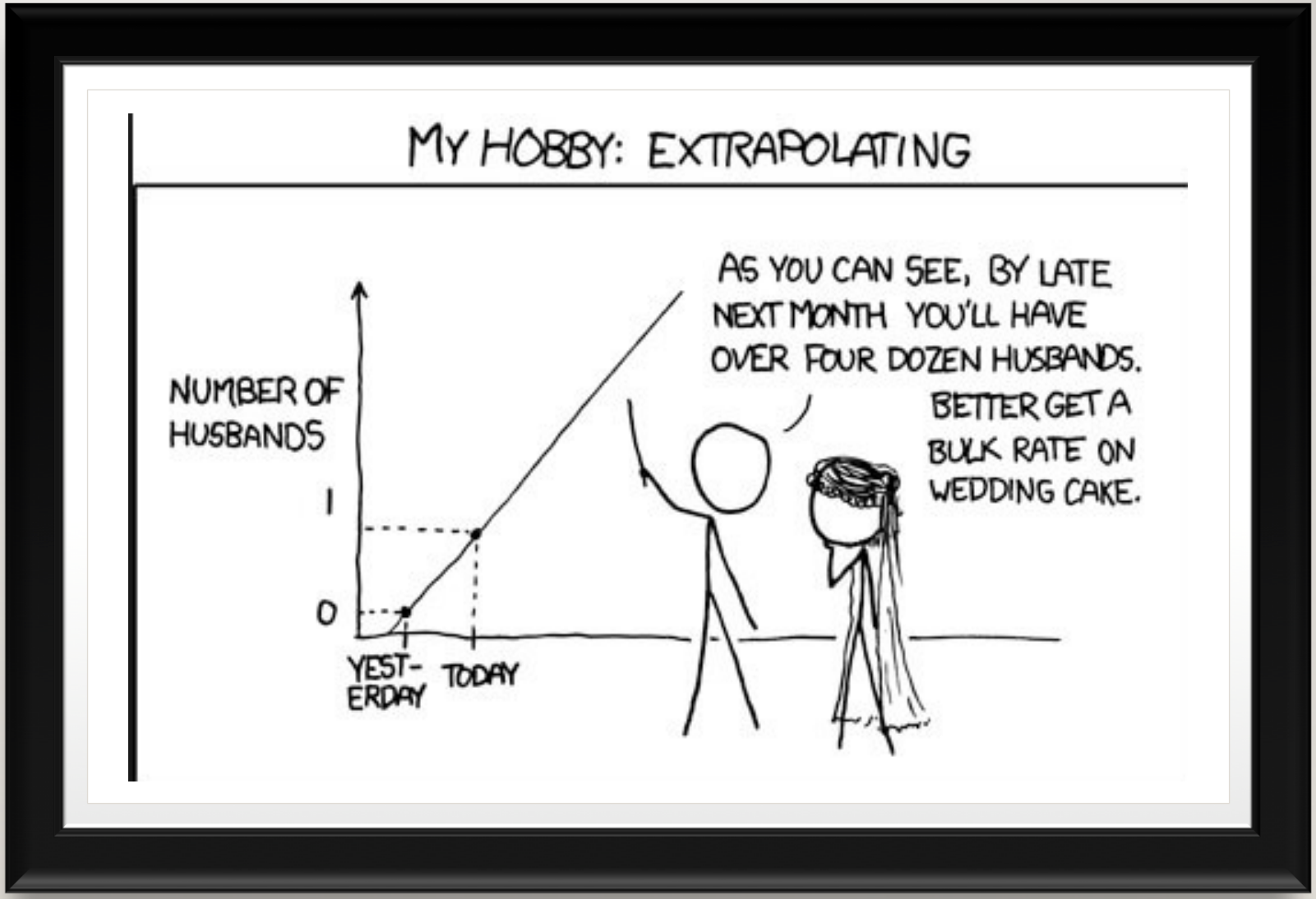
LESSON 12: CORRELATION & REGRESSION

CHRIS QI



OUTLINE

- scatter plots
- correlation
- best fit line
- regression to mean
- least squared estimates

- interpretation of linear model
- residuals
- prediction
- goodness of fit

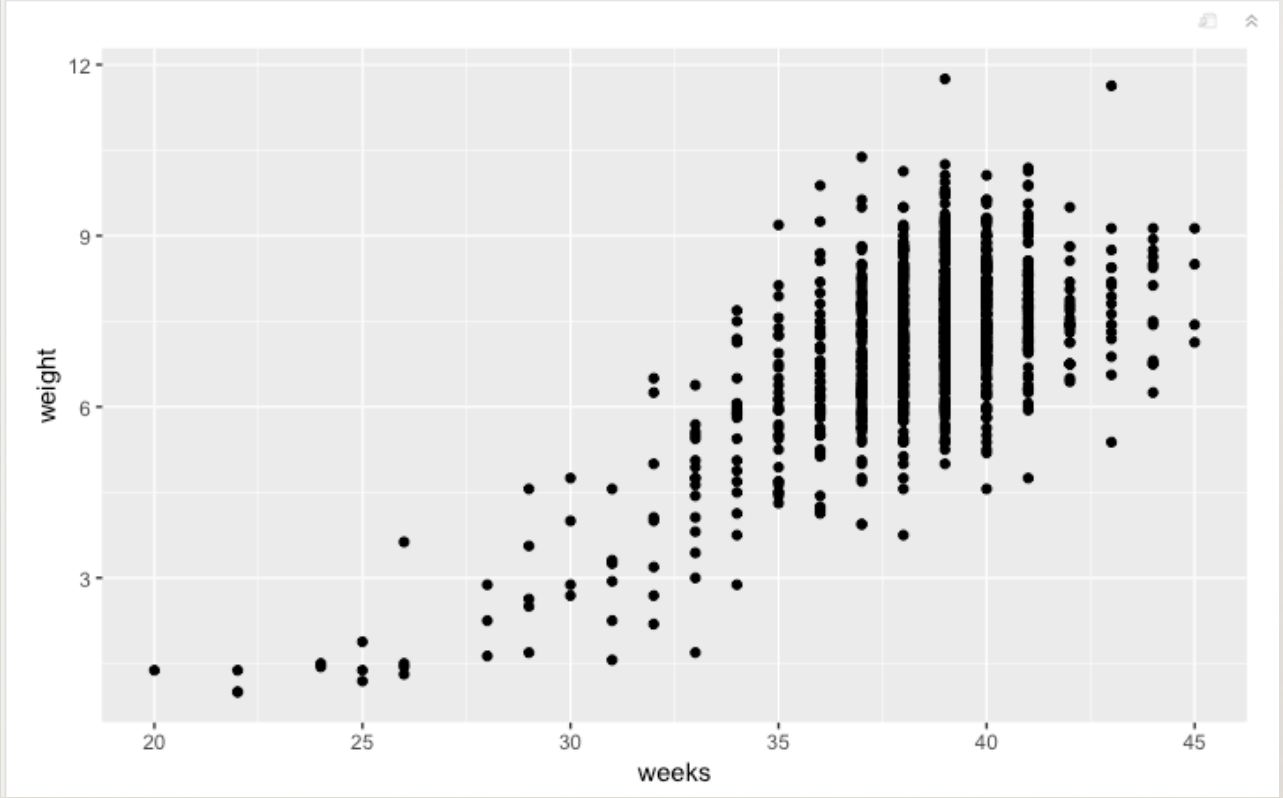
- case study

双变量关系及建模

- 两个变量都是数值型（numeric）
- 应变量（response, dependent variable）
- 解释变量（explanatory）
  - 你认为与应变量有关的东西
  - 也可以叫做，independent variable, predictor

散点图-双变量关系的图形化表示

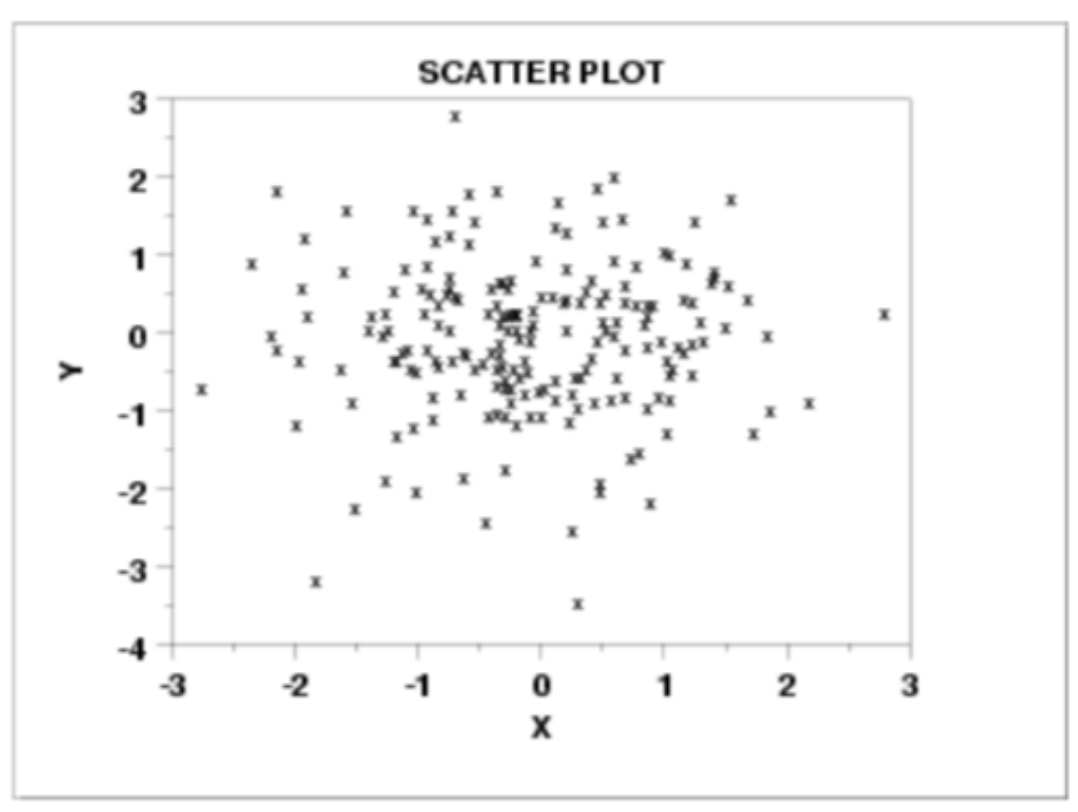
```
ggplot(ncbirths, aes(x=weeks,y=weight))+
geom_point()
```

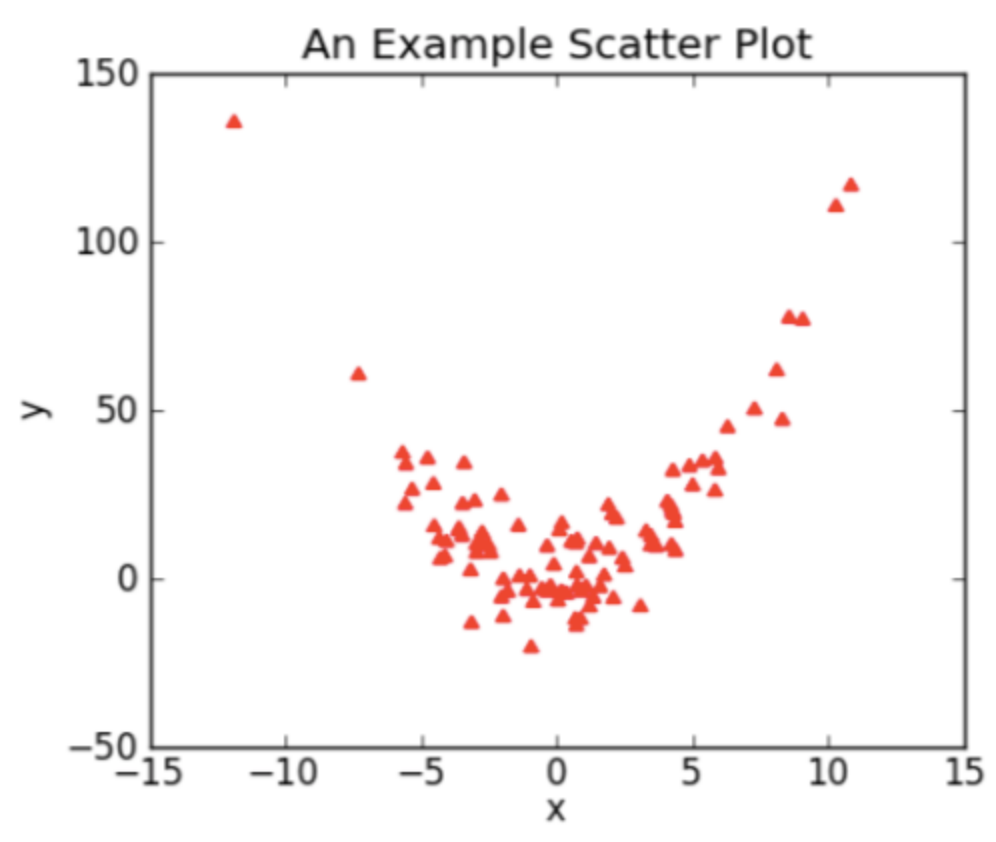


双变量关系的特征

- 形式（线性，非线性）
- 方向（正，负）
- 强度
- 异常值


Sign Legibility Distance (feet)
600
500
400
300
15
25
35
45
55
65
75
85
Driver Age (years)


SCATTER PLOT
Y
X
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3


An Example Scatter Plot
150
100
50
0
−50
y
−15
−10
−5
0
5
10
15
x

相关性
CORRELATION

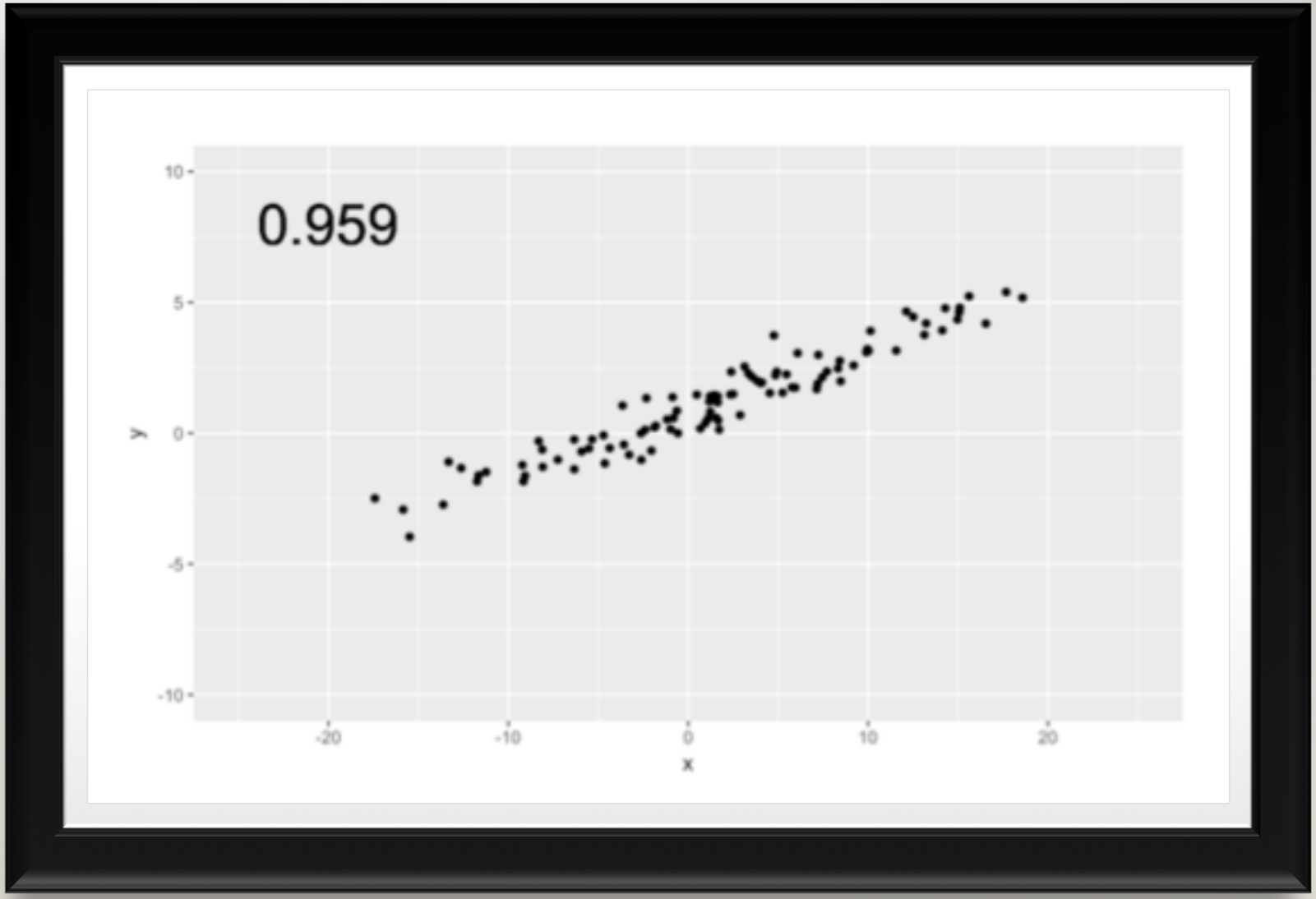
两个变量间的关系到底有多强？

[-1,1]

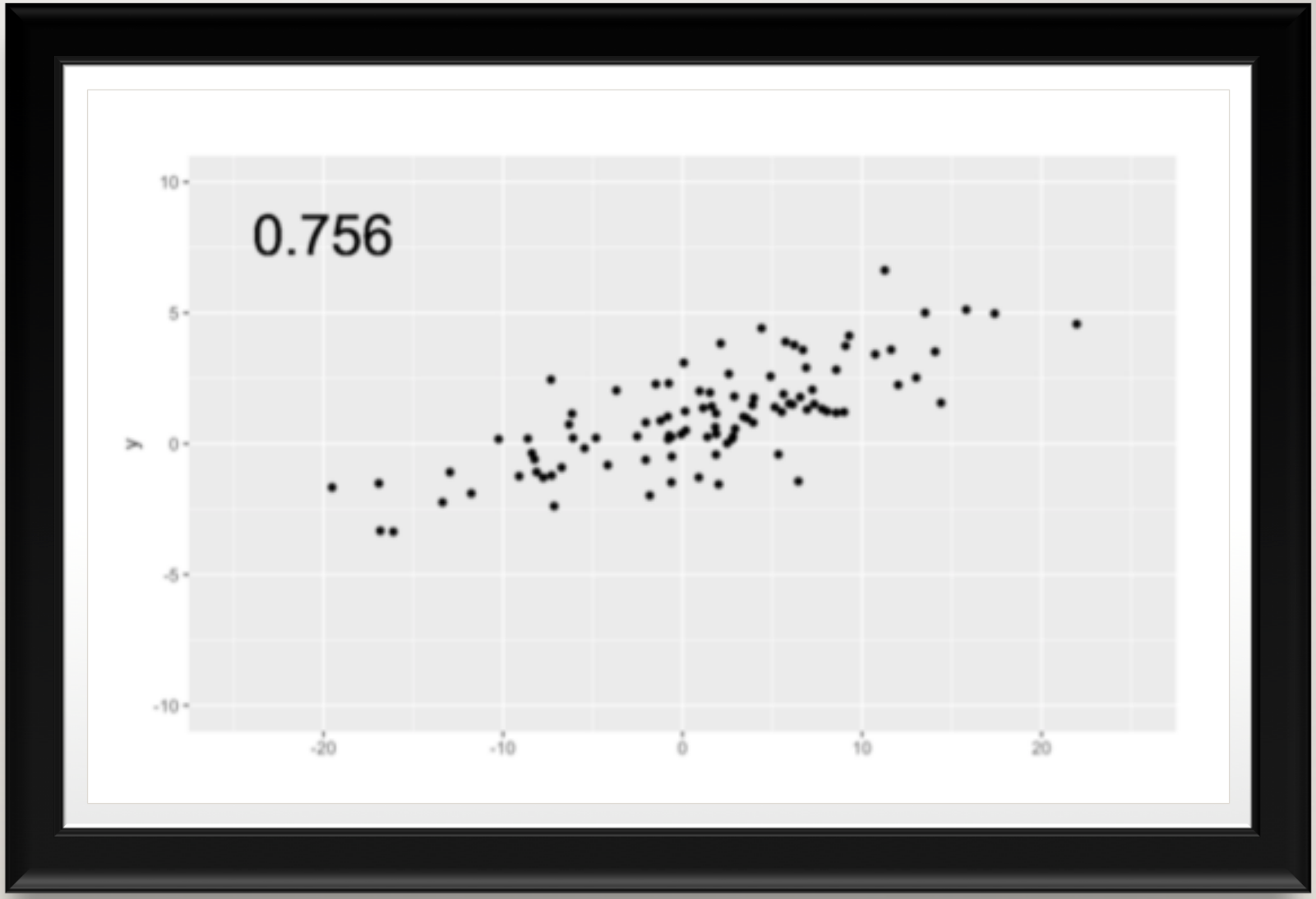
符号正负表示方向

绝对值表示强度

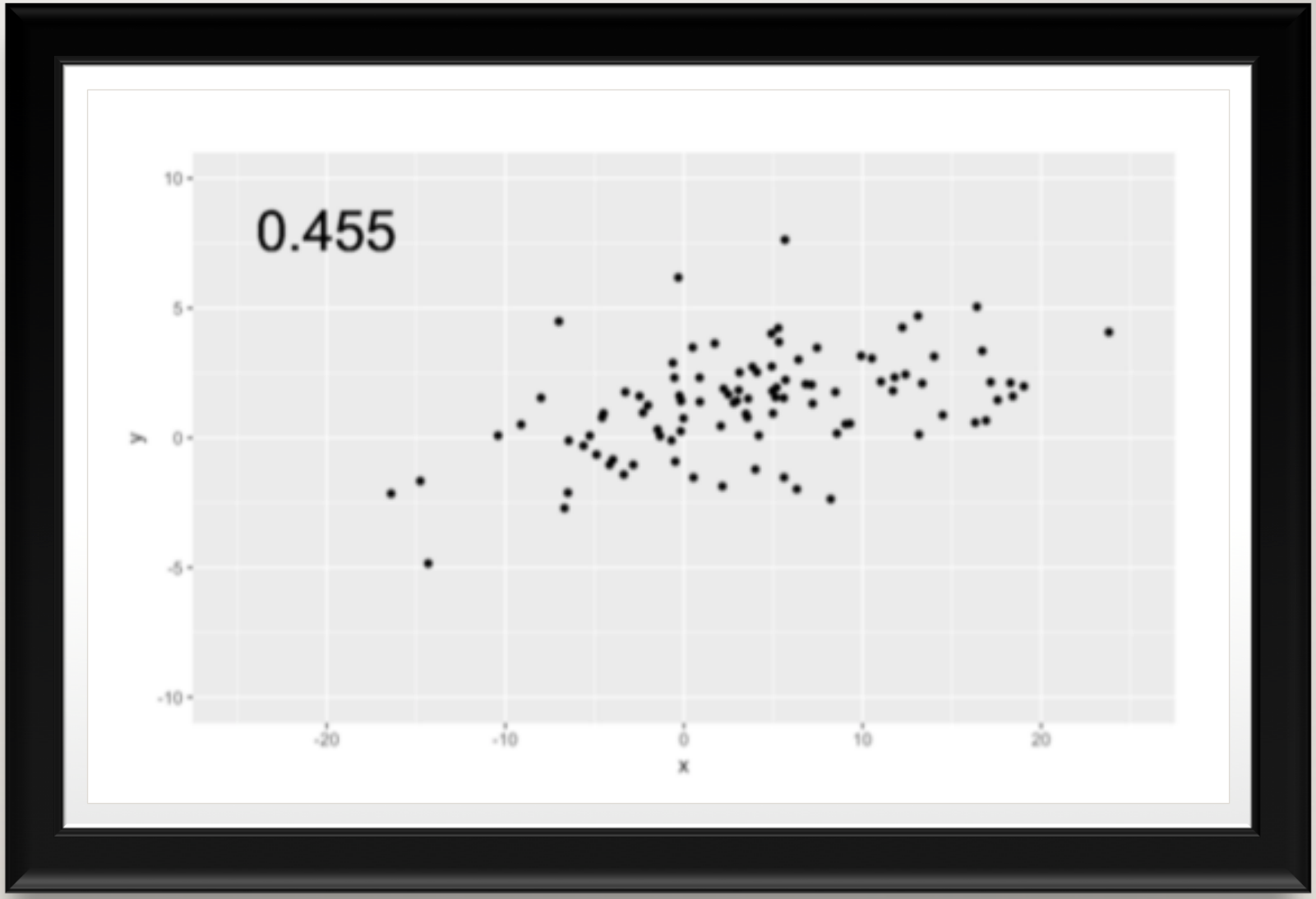
接近完美相关



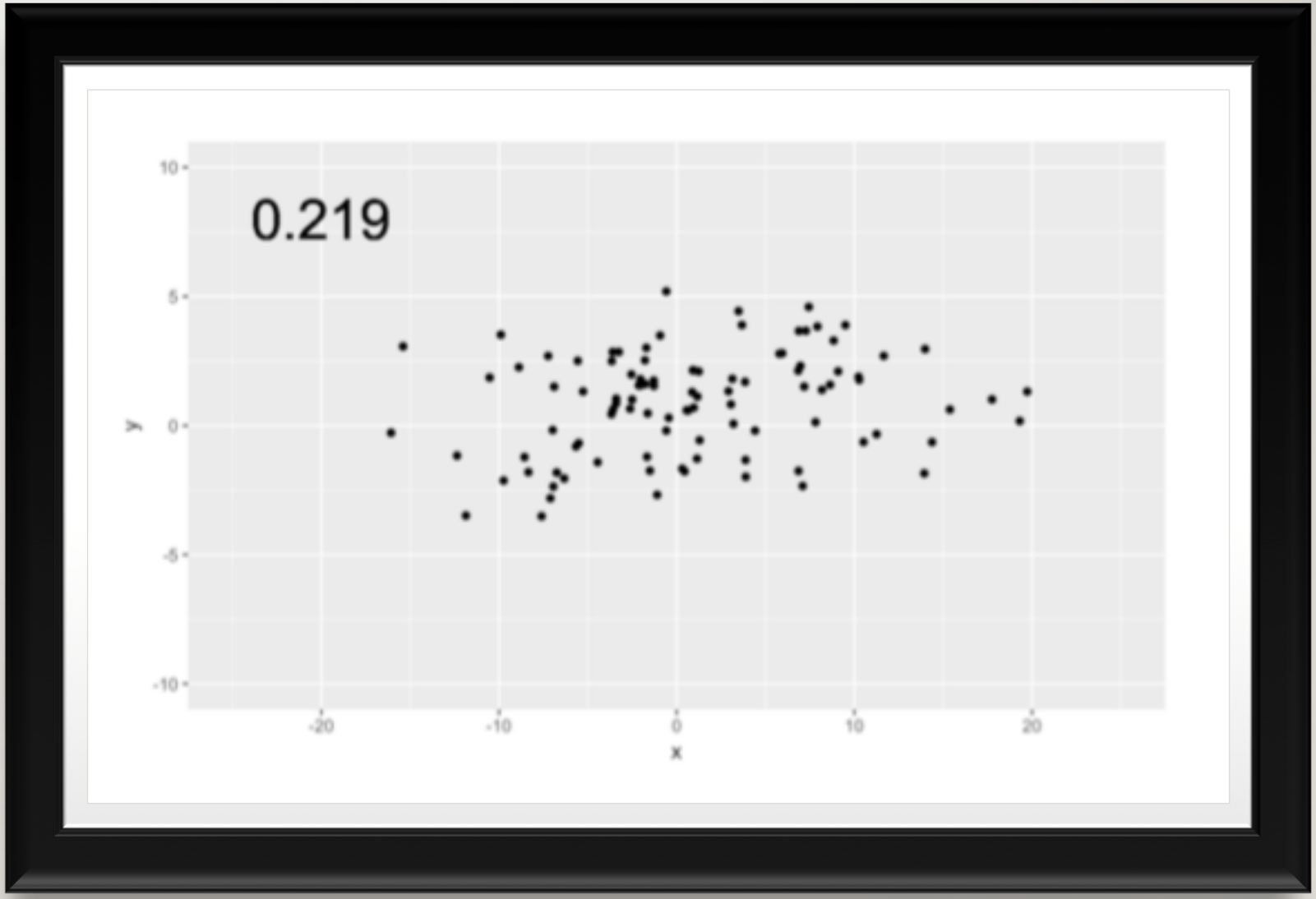
强相关



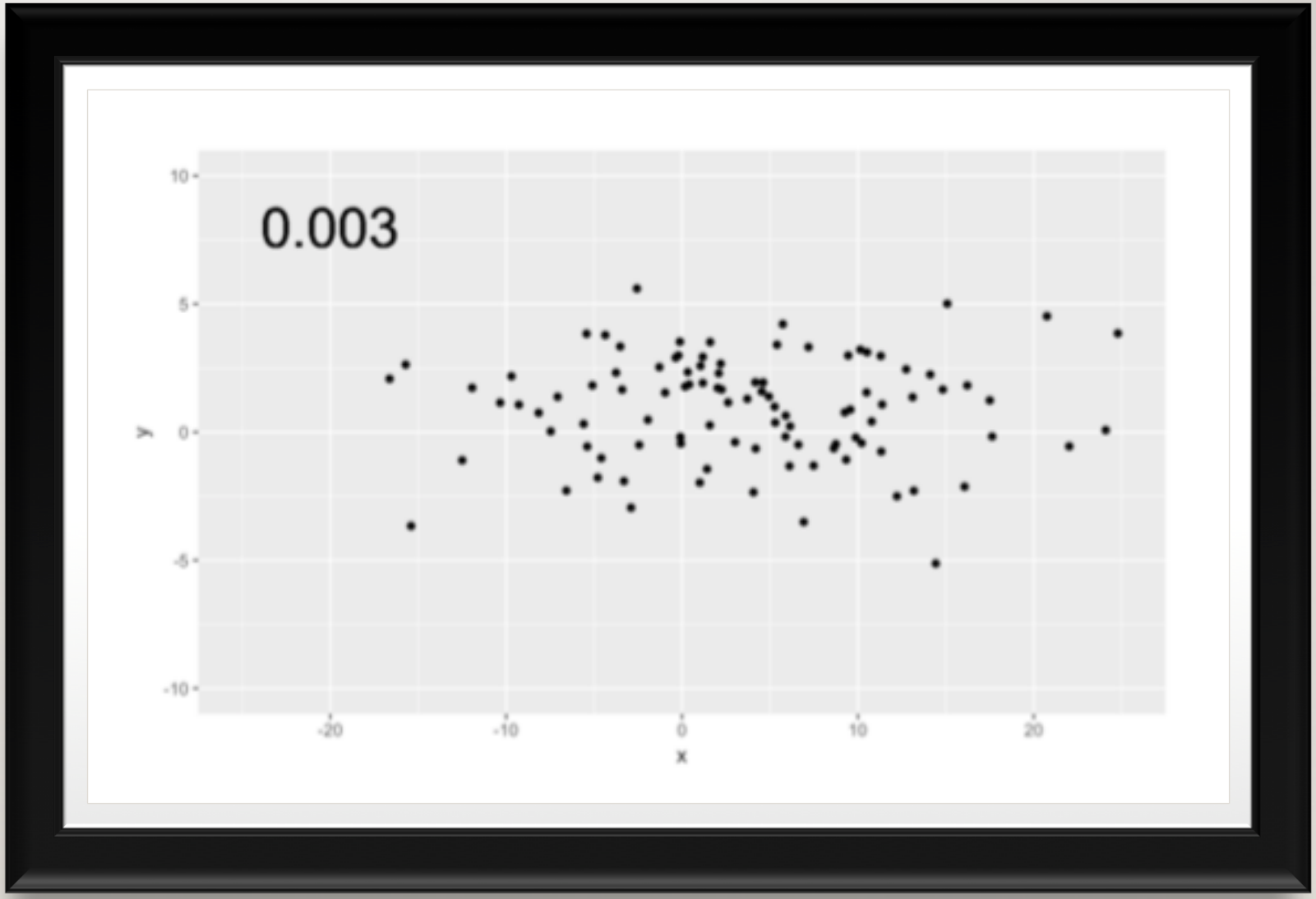
一般



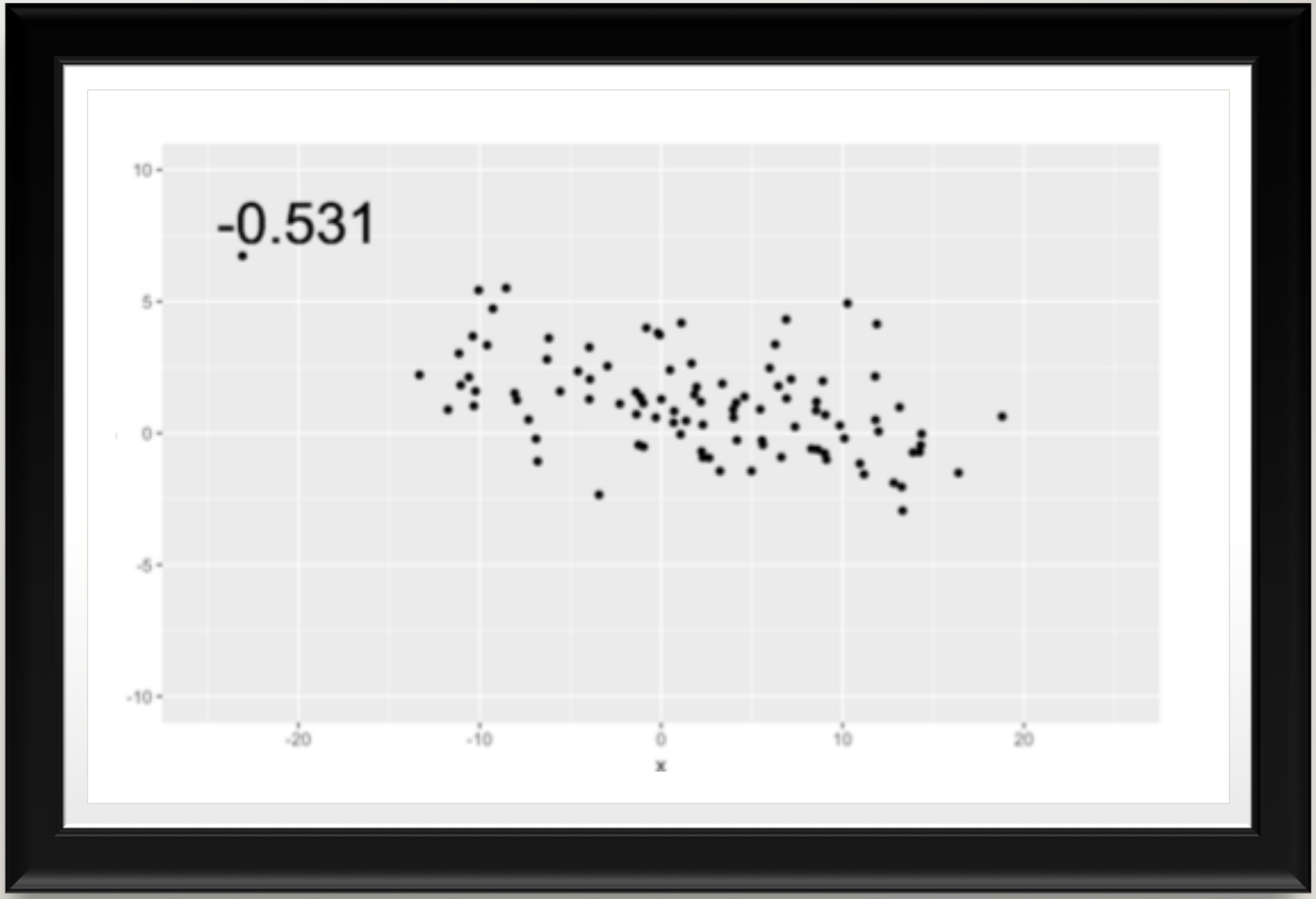
弱相关



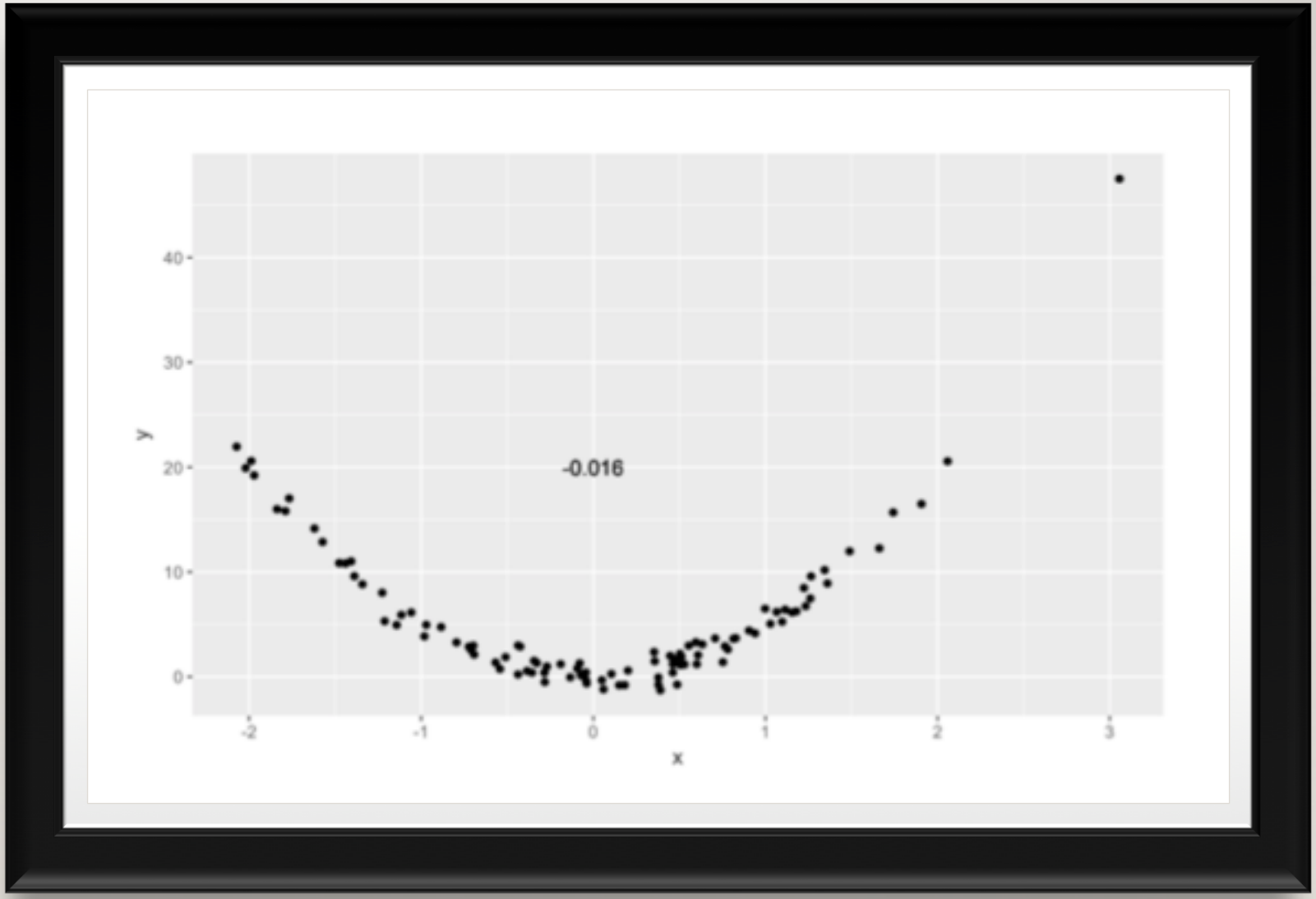
不相关



负相关



非线性



相关系数计算公式

$$r(x,y) = \frac{Cov(x,y)}{\sqrt{SXX \cdot SYY}}$$

相关系数计算公式

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

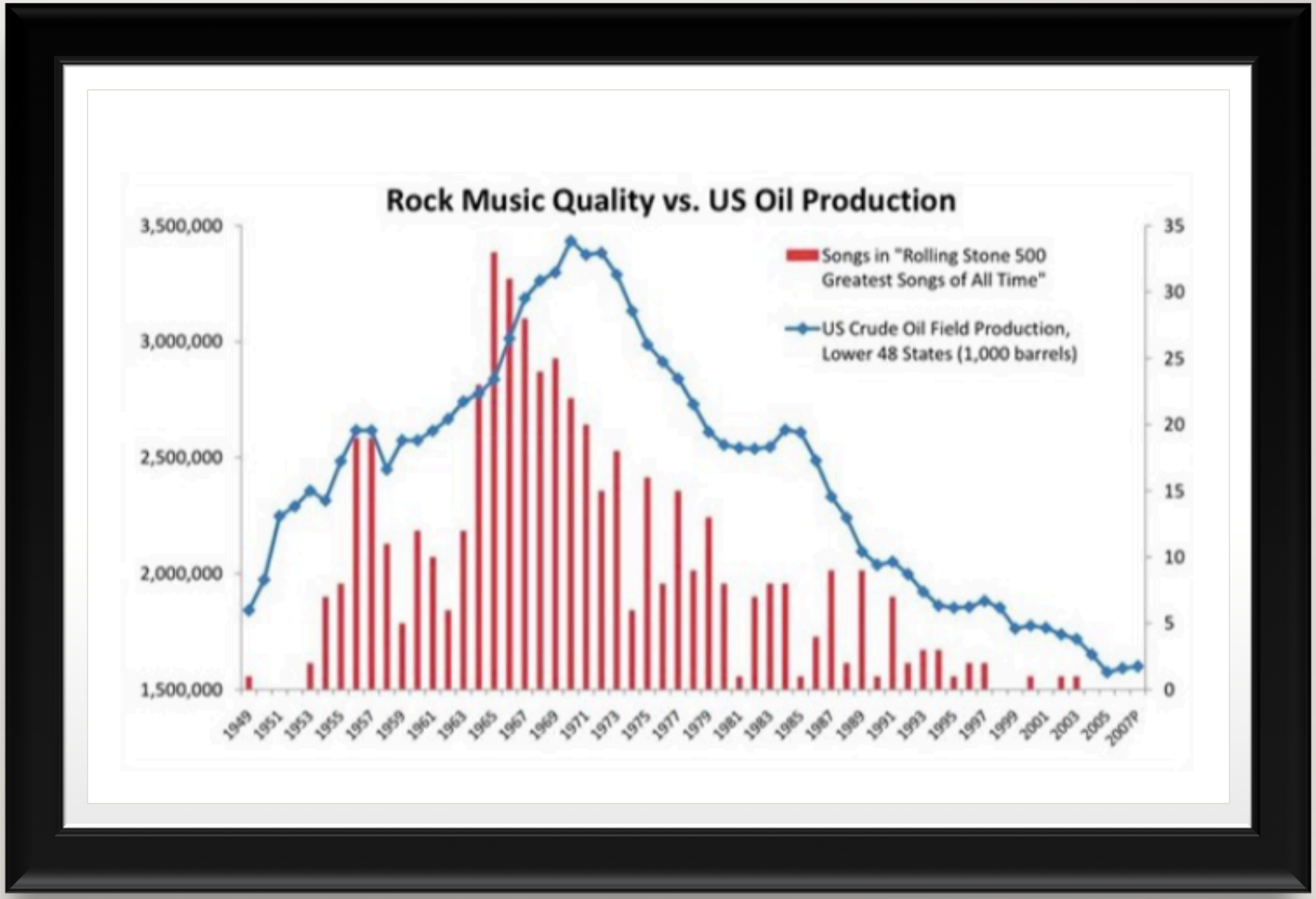
相关性的理解

相关性不等于因果关系 (癌症与吸烟)

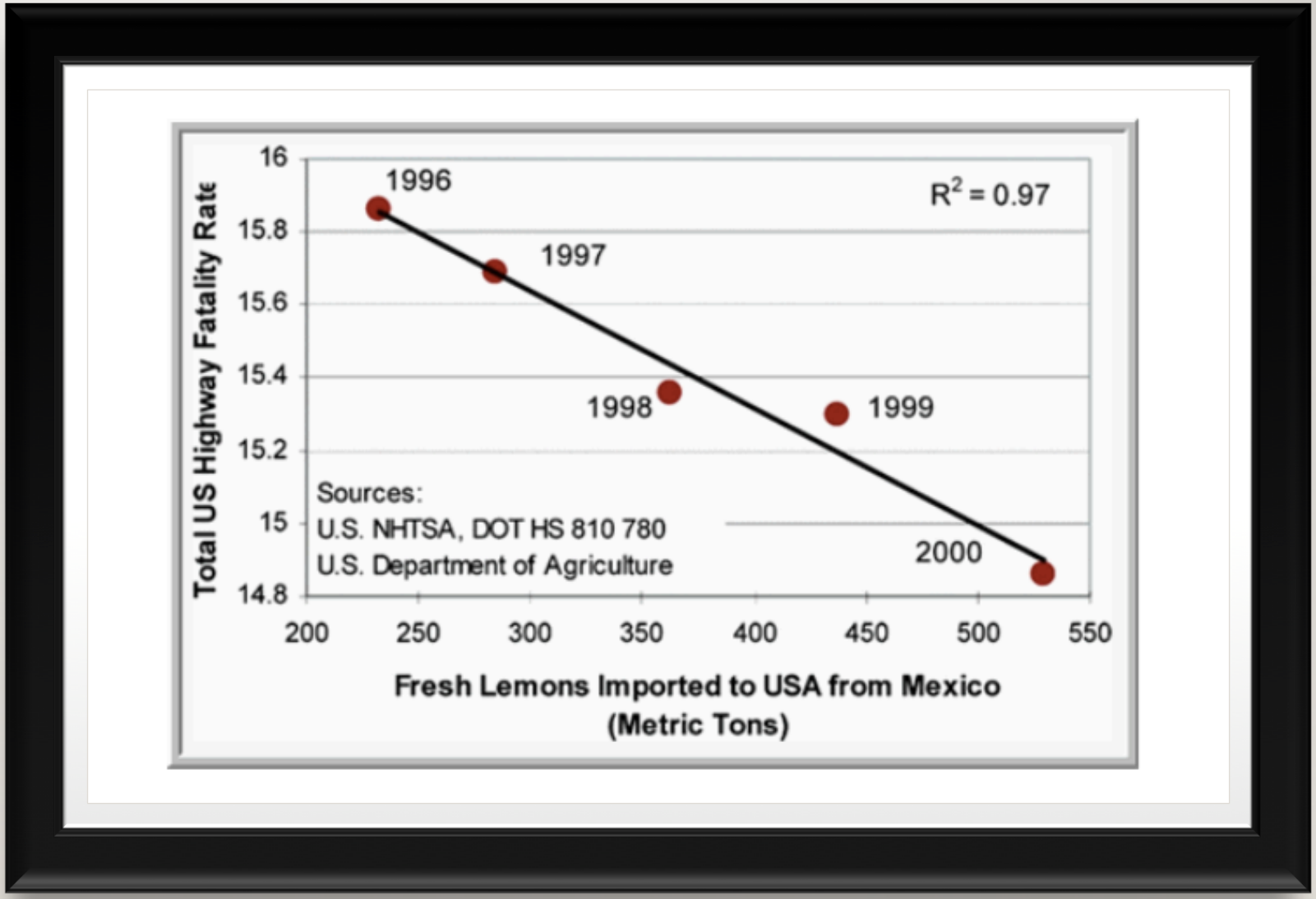
尤其注意因为时间联系在一起的数据，他们的相关性可能毫无意义

尼古拉的电影与游泳池溺亡有关系吗？

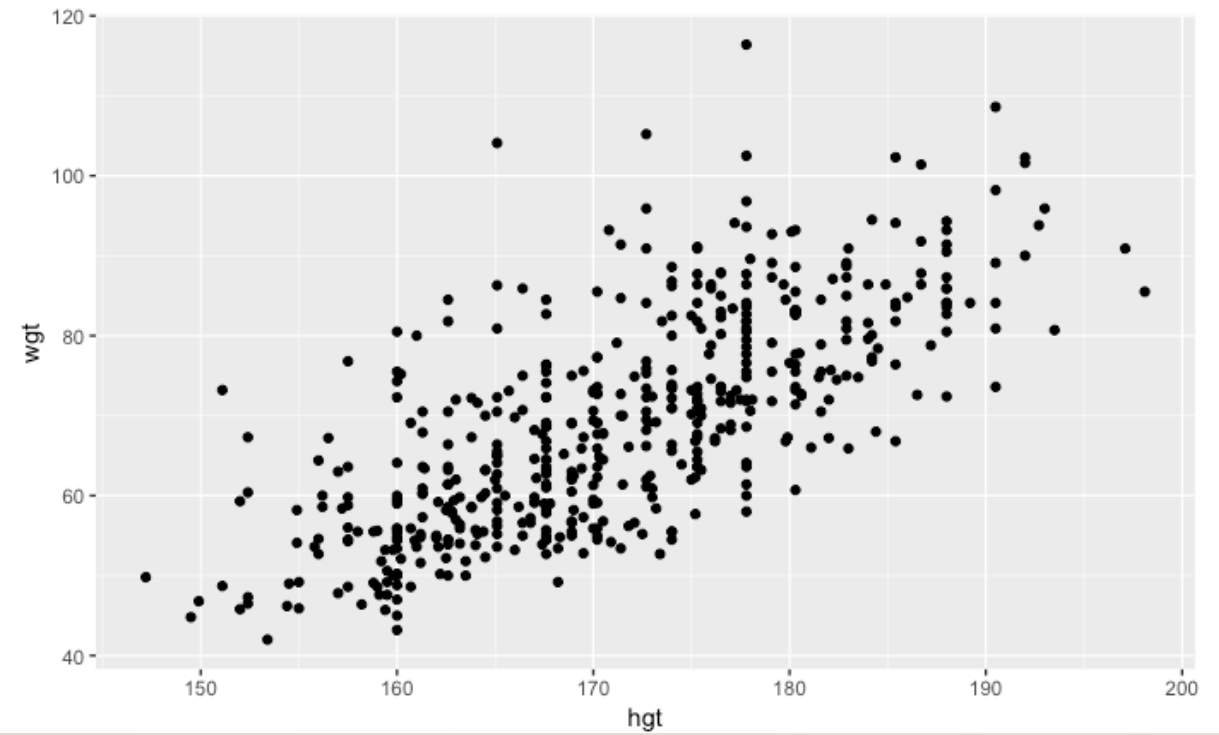
摇滚乐与美国石油产量有关系吗？



高速路死亡率与鲜柠檬进口量有关系吗？

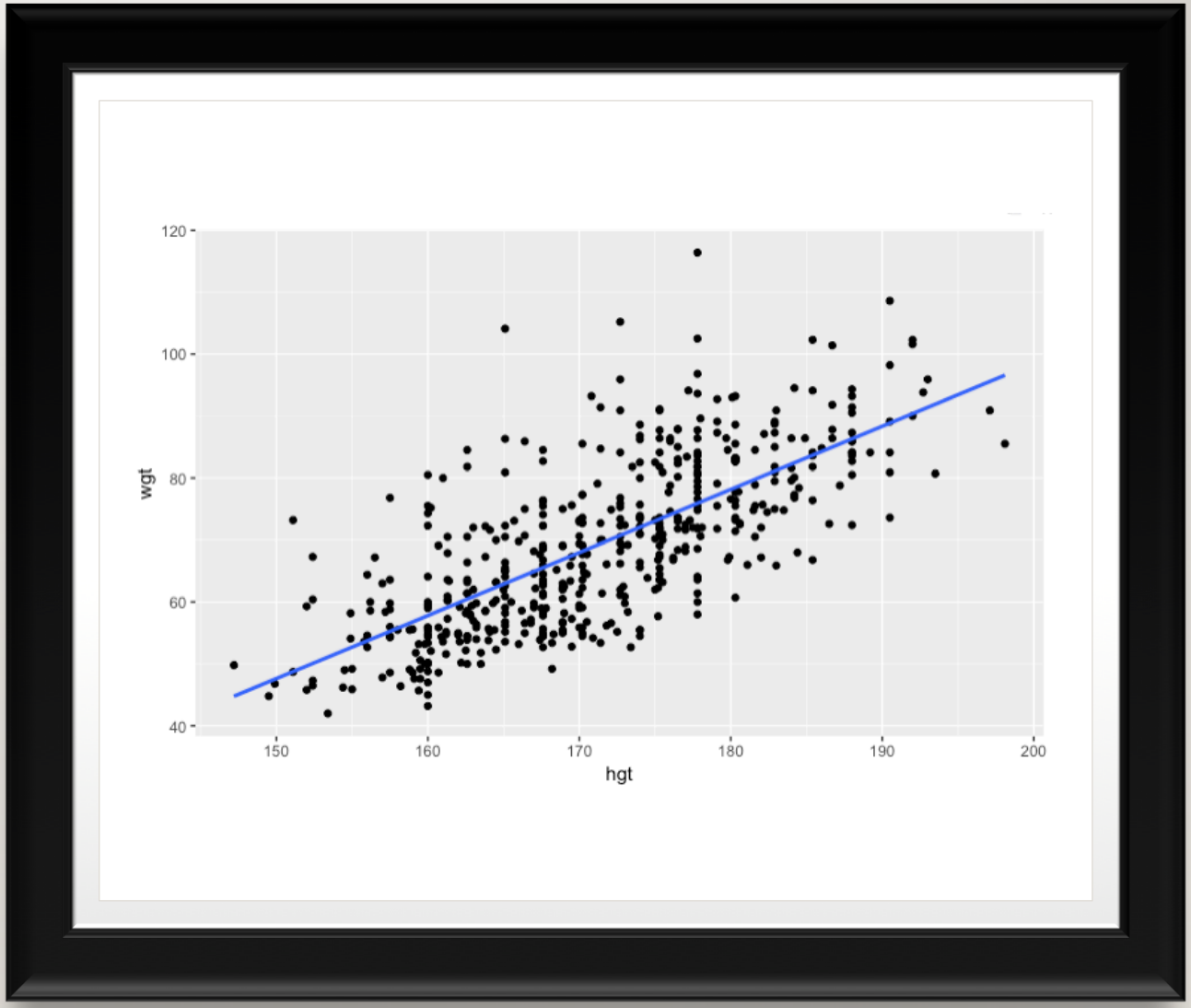


基本回归模型

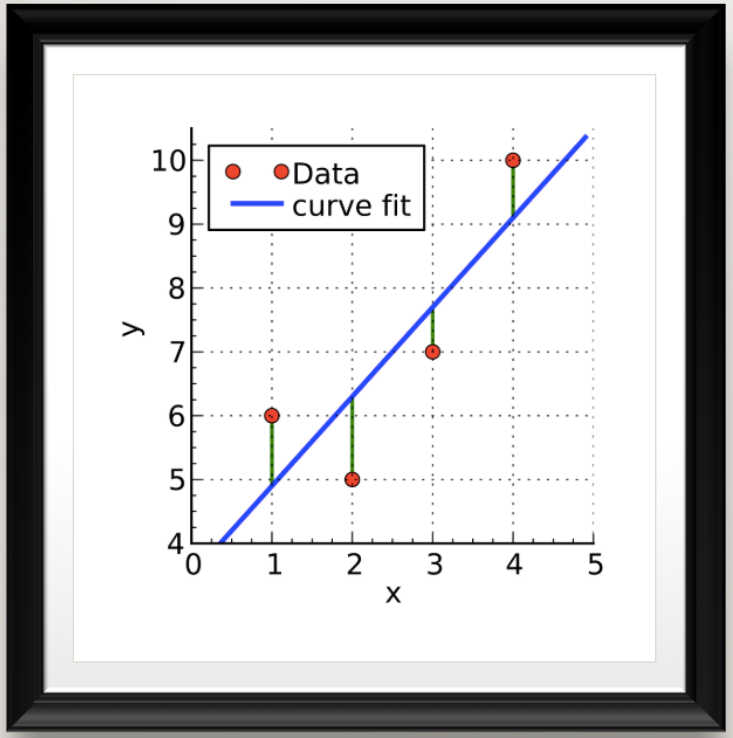


基本回归模型

- $Y=a+b \cdot X$
- 为什么这一条线最合适？



最小二乘法（LEAST SQUARES METHOD）



- 简单地说，最小二乘的思想就是要使得**观测点和估计点的距离的平方和达到最小**.这里的“二乘”指的是用平方来度量观测点与估计点的远近（在古汉语中“平方”称为“二乘”），“最小”指的是参数的估计值要保证各个观测点与估计点的距离的平方和达到最小。

一般的统计模型

- 应变量=f(解释变量)+噪音
- 应变量=截距+斜率*解释变量+噪音

回归模型

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

FITTED VALUE

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

残差（RESIDUAL）

$$e = Y - \hat{Y}$$

拟合的过程：

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

$$e = Y - \hat{Y}$$

- 有n个样本量，也就是n个数据对 （x_i,y_i）
- 找到 截距和斜率，使得残差平方和最小

$$\sum_{i=1}^{n} e_i^2$$

关键概念

- Y-hat 是给定x值时，y的期望值
- Bea-hats是真实但是未知的beta的估计值
- 残差：residual, error, noise

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

$$Y_i = \beta_0 + \beta_1 X + \epsilon$$ 最小二乘法求线性回归系数

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$f = \sum e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial f}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\frac{\partial f}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i$$

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$


别
总
止
!
,

$$Y_i = \beta_0 + \beta_1 X + \epsilon$$ 最小二乘法求线性回归系数

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$f = \sum e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial f}{\partial \hat{\beta}_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\frac{\partial f}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i$$

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

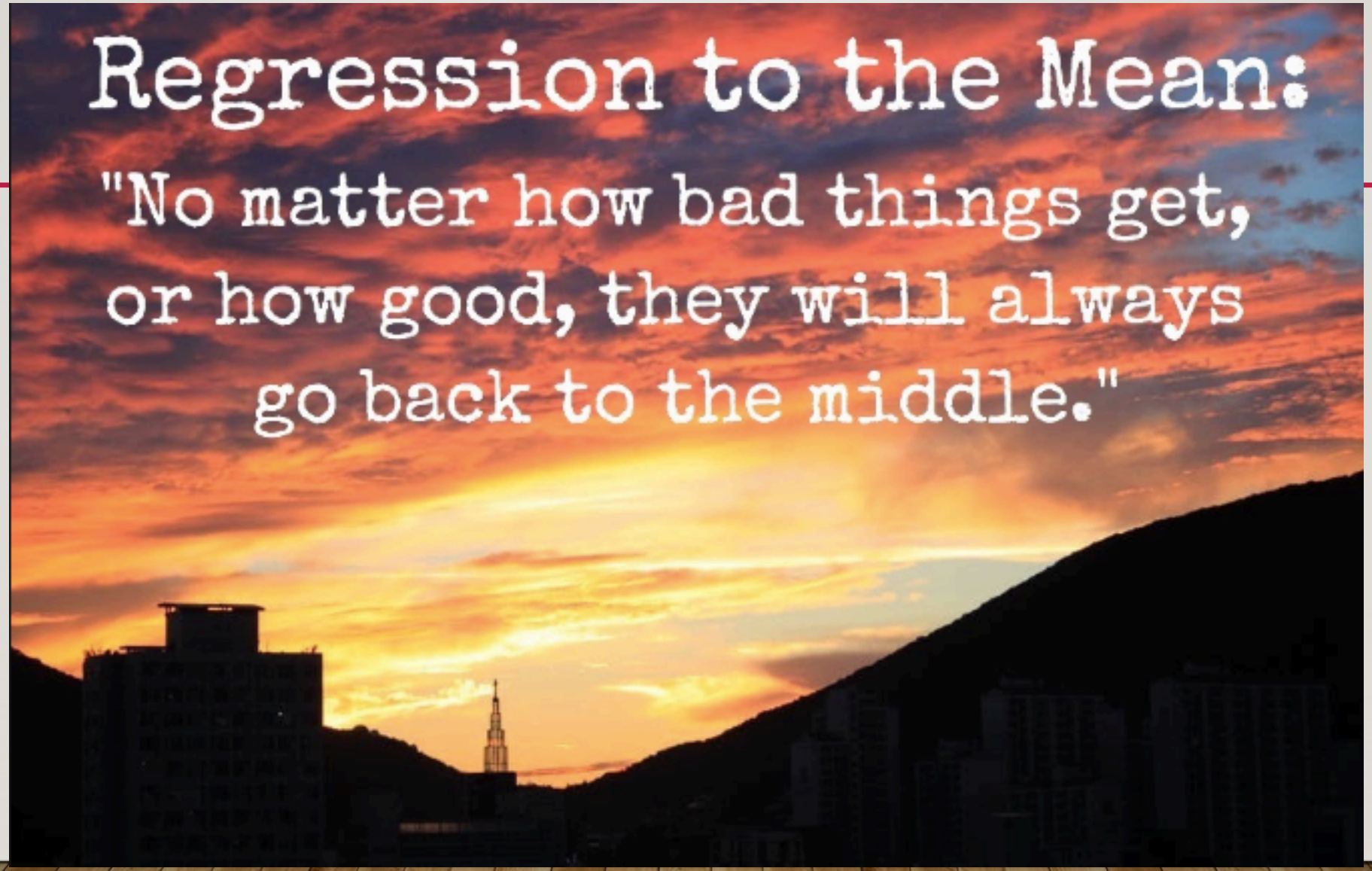
https://my.oschina.net/keyven/blog/526010

最小二乘法的特点

- 唯一性
- 简单（相对于非线性）
- 残差之和等于0
- 回归线必过 $(\bar{x}, \bar{y})$

轶事：回归的由来

- *Regression to the mean is a concept attributed to Sir Francis Galton. The basic idea is that extreme random observations will tend to be less extreme upon a second trial. This is simply due to chance alone. Note that "regression to the mean" and "linear regression" are not the same thing.*
- 孩子的身高和父母的身高：高个子的父母一般生的孩子个子也高，但是孩子一般没有父母那么高，孩子的身高会趋向于平均数。
- 这就是我们没有看到乔丹的儿子能够和乔丹一样成为飞人的原因


Regression to the Mean:
"No matter how bad things get,
or how good, they will always
go back to the middle."