Statistical analysis and data mining complimentary

Daria Bystrova

e-mail: daria.bystrova@inria.fr

https://sites.google.com/view/dariabystrova/home

Information

- ***Time***: 11.30 - 13H room F115
- ***Duration***: 11 weeks
- ***Content***: additional theory, exercises and lab works for the main course.
- ***Grading***: course will be graded by the practical works. (1)
- Books:
    - *The Elements of Statistical Learning*
    - *An Introduction to Statistical Learning*

Multiple linear regression

Model: We have real-valued output Y and input vector $X = (X_1, \ldots, X_p)$ with p predictors:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

- $\beta = (\beta_\mathbf{0}, \ldots, \beta_\mathbf{p})$ unknown constants called coefficients or parameters
- β_0 is called intercept
- ϵ random error term
- If we estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ then

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p,$$

$\hat{y}$ is prediction of Y on the basis of $X = (x_1, \ldots, x_p)$

Simple linear regression model

We assume a model:

slope

predictor

$$Y = \beta_0 + \beta_1 X + \epsilon$$

dependent variable

intercept

error

How do we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$?

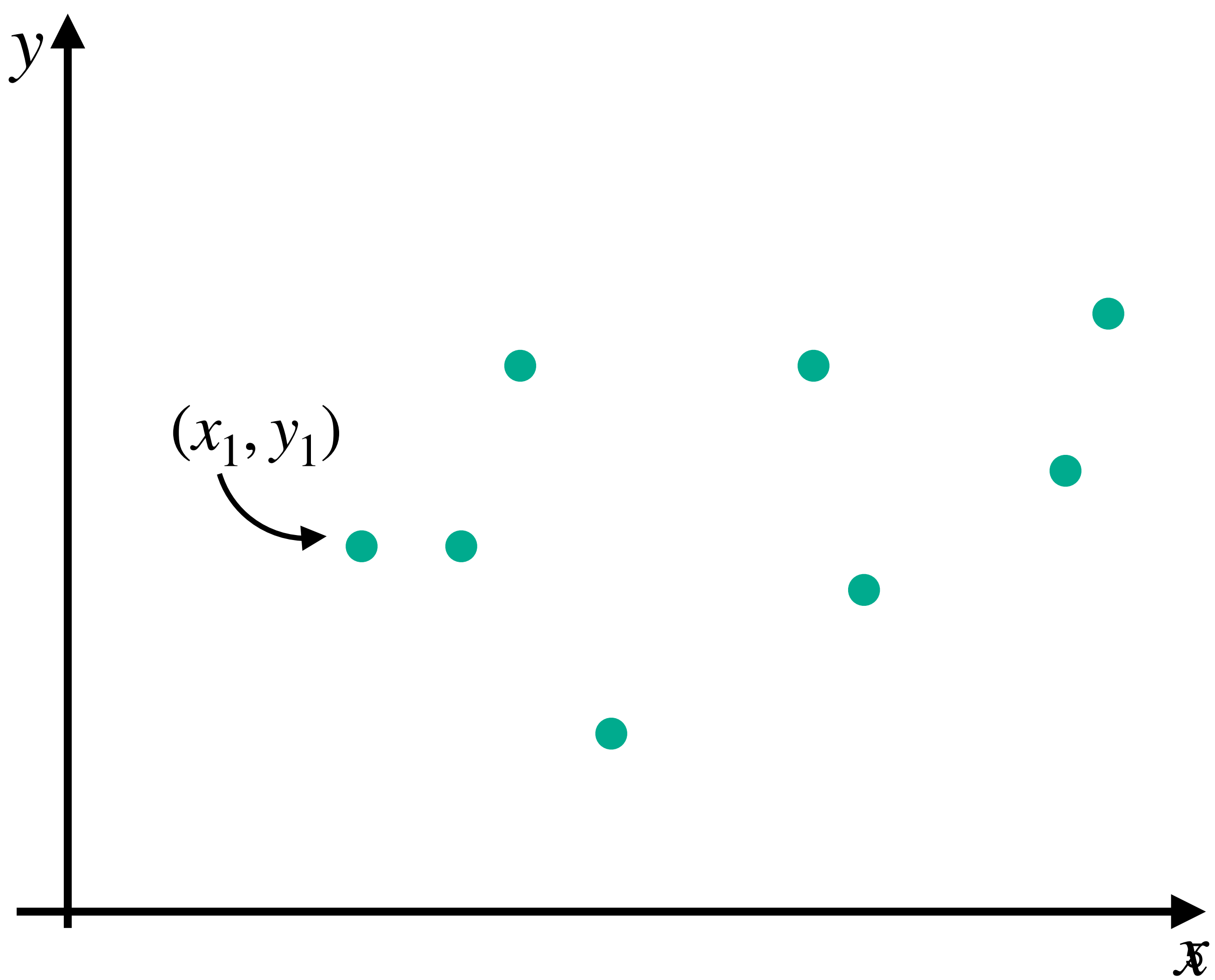
Simple linear regression model

Our model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$:

We have n independent observations: $\{(x_i, y_i)\}_{i=1}^{n}$



$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i \quad i = 1, \ldots, n$$

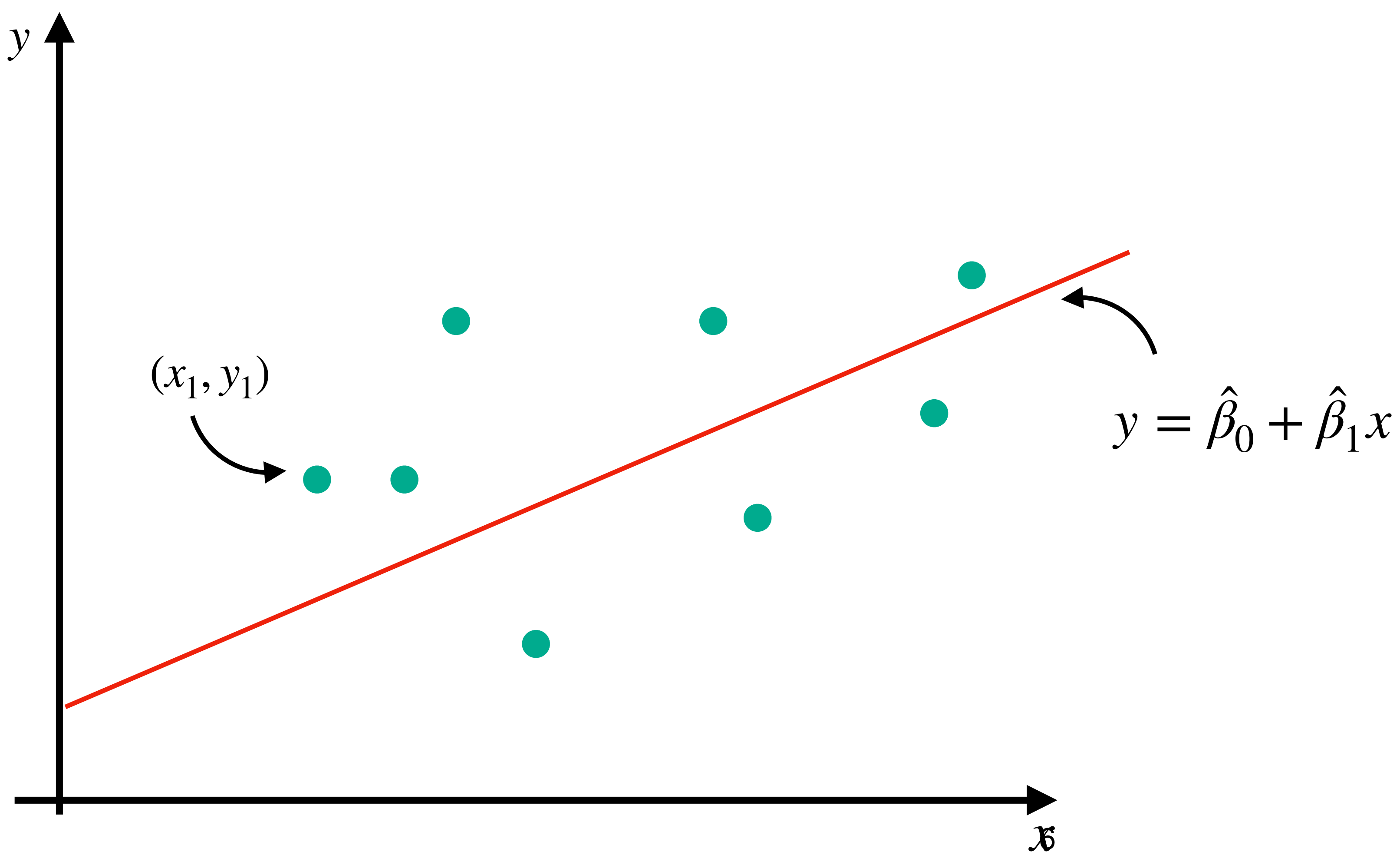
Simple linear regression model

Our model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$:

We have n independent observations: $\{(x_i, y_i)\}_{i=1}^{n}$



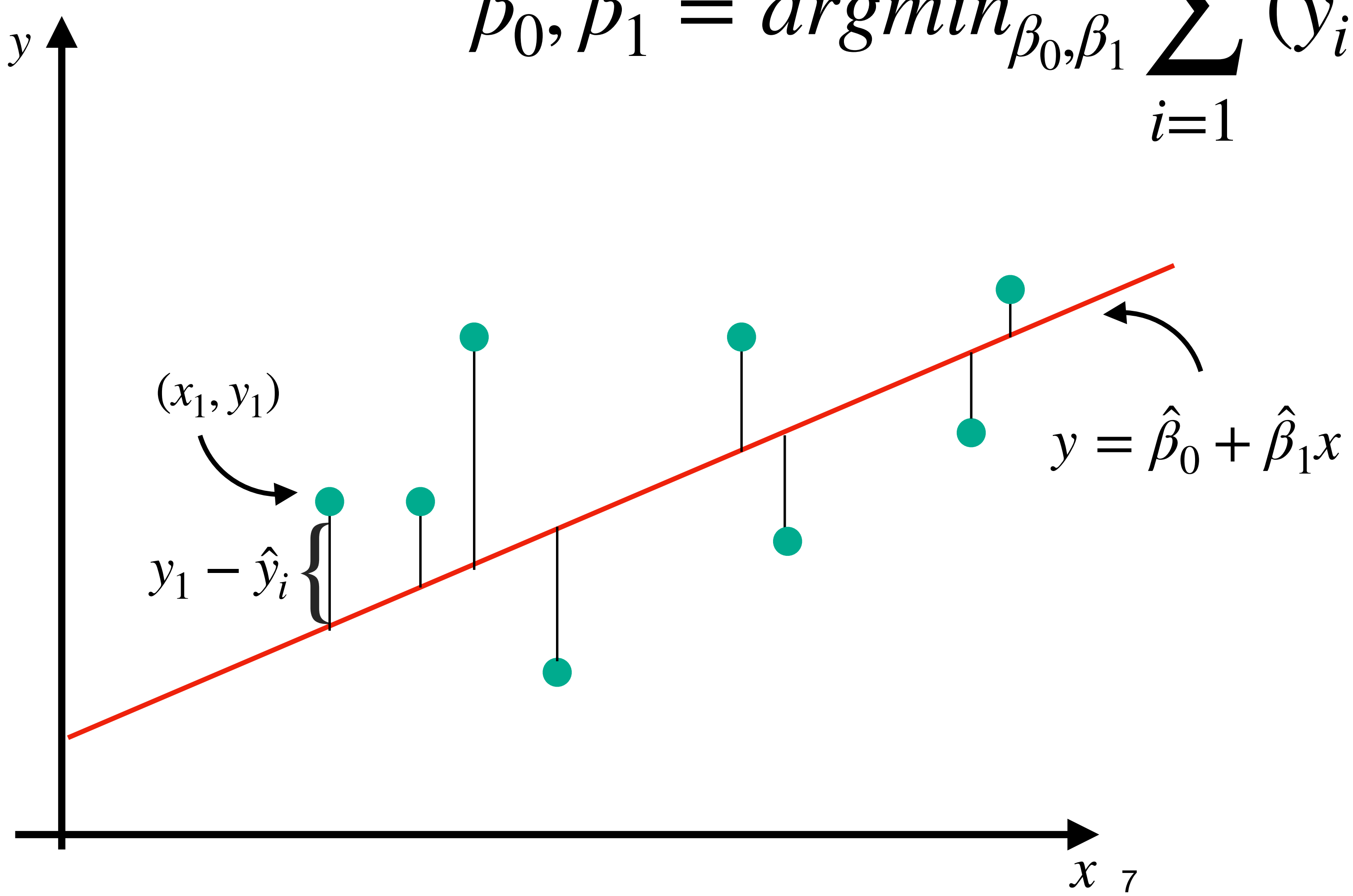
Least squares estimators

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad i = 1, \ldots, n$$

We define residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{\beta}_0, \hat{\beta}_1 = argmin_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Ex.1. Find expressions for $\hat{\beta}_0, \hat{\beta}_1$?

Least square estimators

$$\hat{\beta}_0, \hat{\beta}_1 = argmin_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\nabla RSS = (\frac{\partial RSS}{\partial \beta_0}, \frac{\partial RSS}{\partial \beta_1}), \nabla RSS = (0,0)$$

$$\begin{cases} \dfrac{\partial RSS}{\partial \beta_0} = -2\Big[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\Big] = 0 \\ \dfrac{\partial RSS}{\partial \beta_1} = -2\Big[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i\Big] = 0 \end{cases}$$

Least square estimators

$$\hat{\beta}_0, \hat{\beta}_1 = argmin_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\nabla RSS = (\frac{\partial RSS}{\partial \beta_0}, \frac{\partial RSS}{\partial \beta_1}), \nabla RSS = (0,0)$$

$$\begin{cases} \dfrac{\partial RSS}{\partial \beta_0} = -2\Big[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\Big] \implies \beta_0 = \dfrac{1}{n}\sum_{i=1}^{n} y_i. - \beta_1 \dfrac{1}{n} \sum_{i=1}^{n} x_i \\ \dfrac{\partial RSS}{\partial \beta_1} = -2\Big[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i\Big] = 0 \end{cases}$$

Reminder: $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ sample mean, $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x}_n)^2$ sample variance

$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)$ sample covariance

Least square estimators

$$\hat{\beta}_0, \hat{\beta}_1 = argmin_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i. - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y}_n - \beta_1 \bar{x}_n =$$

$$-2\Big[\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i\Big] = \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2 - [\bar{y}_n - \beta_1 \bar{x}_n]\sum_{i=1}^{n} x_i \implies$$

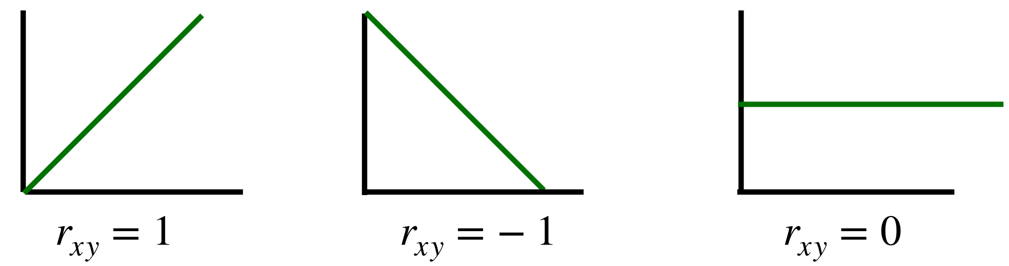
$$\underbrace{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y}_n}_{c_{xy}} - \beta_1(\underbrace{\sum_{i=1}^{n} x_i^2 - \bar{x}_n \sum_{i=1}^{n} x_i}_{S_x^2}) = 0 \implies \boxed{\beta_1 = \frac{c_{xy}}{S_x^2} \text{ slope}}$$

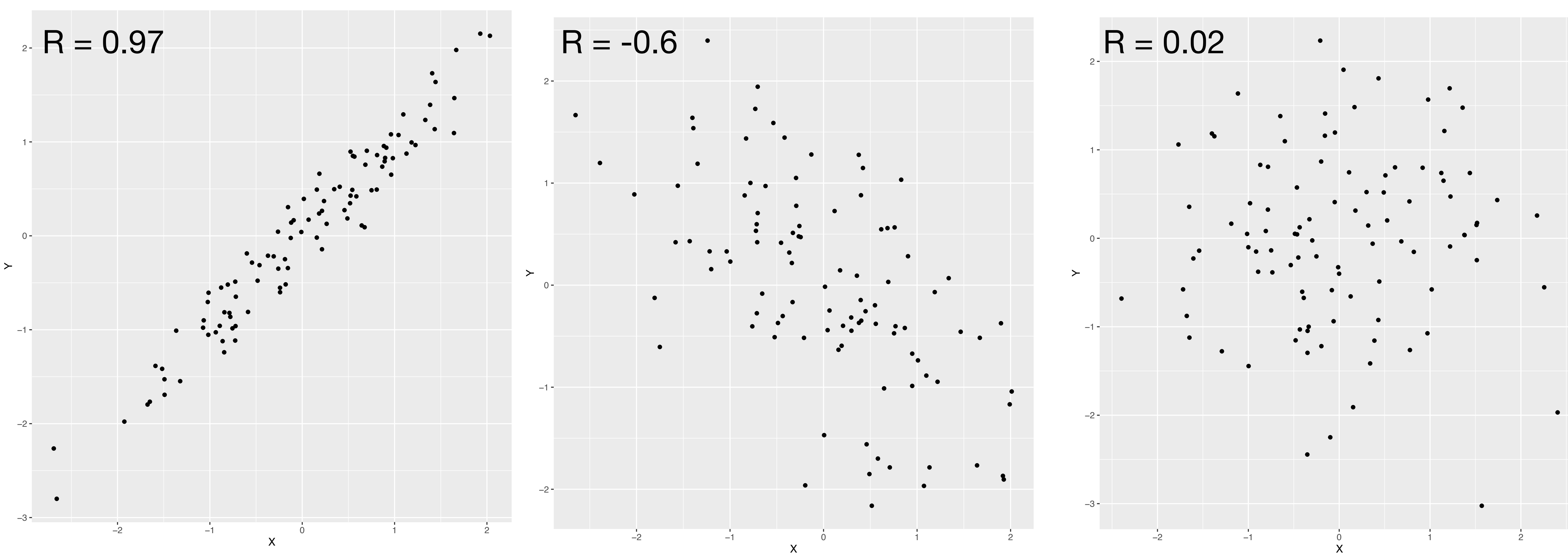
$$\boxed{\text{Intercept } \beta_0 = \bar{y}_n - \frac{S_{xy}}{S_x^2}\bar{x}_n}$$

Correlation

$$r_{xy} = \frac{c_{xy}}{S_x S_y}$$ empirical correlation coefficient

- $r_{xy} \in [-1,1]$
- $r_{xy} = 1$ positive linear dependence
- $r_{xy} = -1$ negative dependence





Correlation

$$r_{xy} = \frac{c_{xy}}{S_x S_y}$$ empirical correlation coefficient

- Correlation coefficient does not change with centering or scaling of the variables

Correlation

$$r_{xy} = \frac{c_{xy}}{S_x S_y}$$ empirical correlation coefficient

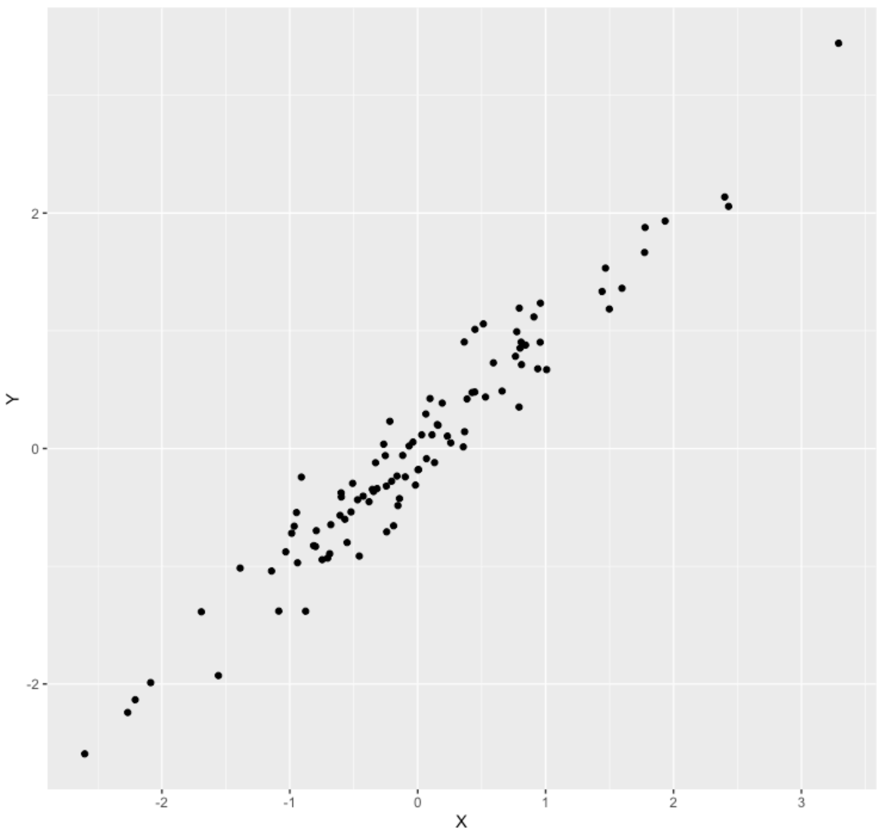
- Correlation coefficient does not change with centering or scaling of the variables
- Correlation is symmetric (correlation of X with Y is the same as X with Y)

Correlation

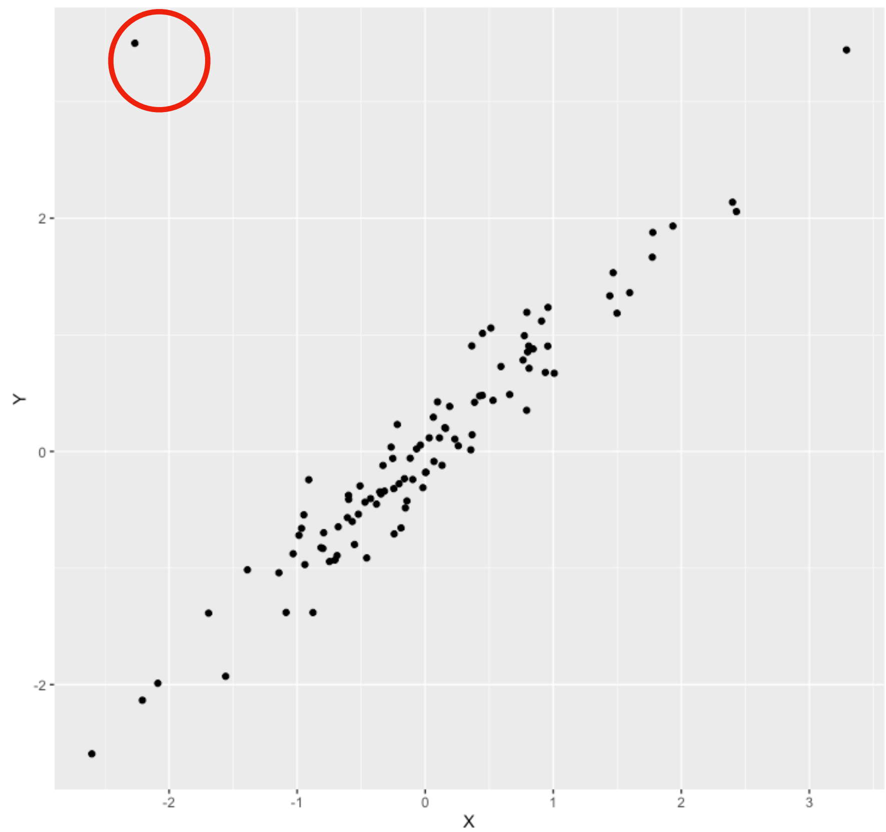
$$r_{xy} = \frac{c_{xy}}{S_x S_y}$$ empirical correlation coefficient

- Correlation coefficient does not change with centering or scaling of the variables
- Correlation is symmetric (correlation of X with Y is the same as X with Y)
- Correlation coefficient is sensitive to outliers

R = 0.97

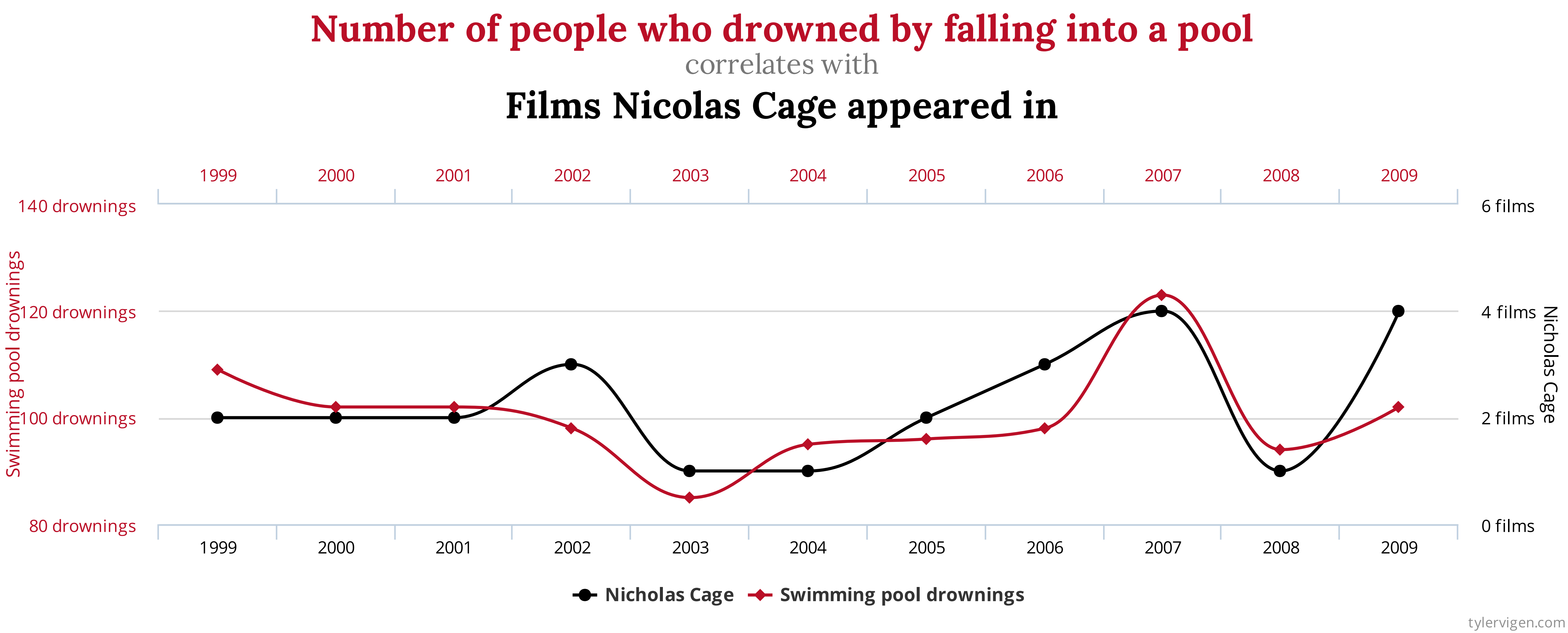


R = 0.81



Correlation

Correlation does not imply causation!



From https://www.tylervigen.com

Question: When RSS is the smallest?

Question: When RSS is the smallest?

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)^2 = \text{[using the derived expressions]} \ \ ?$$

Question: When RSS is the smallest?

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)^2 = \text{[using the derived expressions]} =$$

$$= \sum_{i=1}^{n} (y_i - \overline{y}_n - \frac{c_{xy}}{S_x^2}(x_i - \overline{x}_n))^2 = \; .. = \Big[\sum_{i=1}^{n} (y_i - \overline{y}_i)\Big](1 - \frac{c_{xy}^2}{S_x^2 S_y^2}) =$$

$$= \Big[\sum_{i=1}^{n} (y_i - \overline{y}_i)^2\Big](1 - r_{xy}^2)$$

Answer: When $r_{xy} = \pm 1$

Accuracy of the model

- Residual standard error $RSE = \sqrt{\frac{1}{n-2}RSS}$ absolute measure lack of fit

-

Accuracy of the model

- Residual standard error $RSE = \sqrt{\frac{1}{n-2}RSS}$ absolute measure lack of fit
- R^2 statistic measure proportion of variance explained
    - Residual Sum of Squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$
    - Total Sum of squares $TSS = \sum_{i=1}^{n}(y_i - \overline{y}_n)$
    - $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

Accuracy of the model

- Residual standard error $RSE = \sqrt{\frac{1}{n-2}RSS}$ absolute measure lack of fit

- R^2 statistic measure proportion of variance explained
    - Residual Sum of Squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$
    - Total Sum of squares $TSS = \sum_{i=1}^{n}(y_i - \overline{y}_n)$
    - $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

In case of simple linear regression:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{TSS(1 - r_{xy}^2))}{TSS} = r_{xy}^2$$

Properties of estimators $\hat{\beta}_0$, $\hat{\beta}_1$

$$Y = \beta_0 + \beta_1 x + \epsilon$$

- We assume that random term ϵ has $\mathbb{E}[\epsilon] = 0$
- $\{(x_i, y_i)\}_{i=1}^n$ represent i.i.d random sample of size n
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators:
    - $\mathbb{E}[\hat{\beta}_0] = \beta_0$
    - $\mathbb{E}[\hat{\beta}_1] = \beta_1$
- $Var[\hat{\beta}_1] = \dfrac{Var[\epsilon]}{nS_x^2}$
- $Var[\hat{\beta}_0] = \dfrac{Var[\epsilon]}{n}(1 + \dfrac{\overline{x}_n}{S_x^2})$