Tutorial 8

Statistical Computation and Analysis

Spring 2023

Outline

- Correlation

- Simple Linear Regression
  - estimating coefficients and error
  - hypothesis testing for slope
  - confidence interval for regression line

- Multivariate Linear Regression

Correlation

Measures for correlation between two variables

- Pittman correlation coefficient

$$Pittman = \sum_{i=1}^{N} x_i y_i$$

- Pearson correlation coefficient

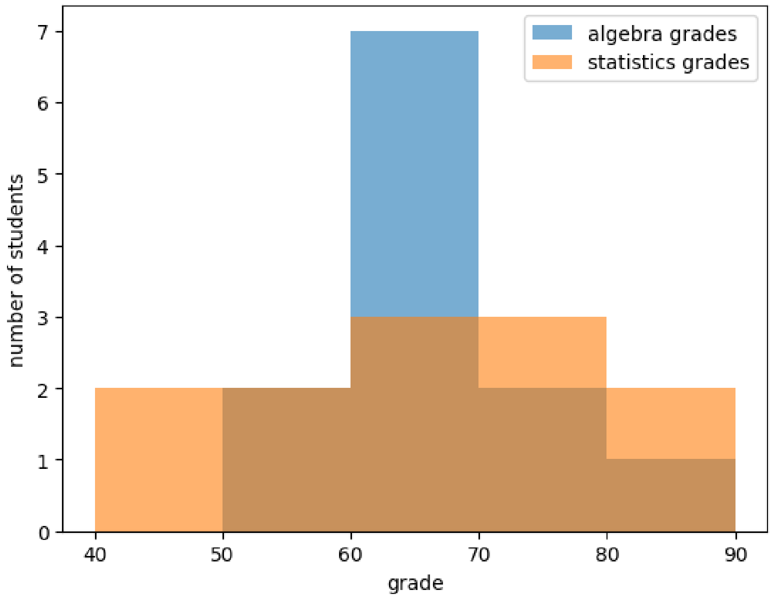
$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Example 1 - correlation

We aim to check if better knowledge in algebra helps to better understand statistics.

We have data (Mardia, Kent and Bibby, 1979) about students grades in algebra and statistics

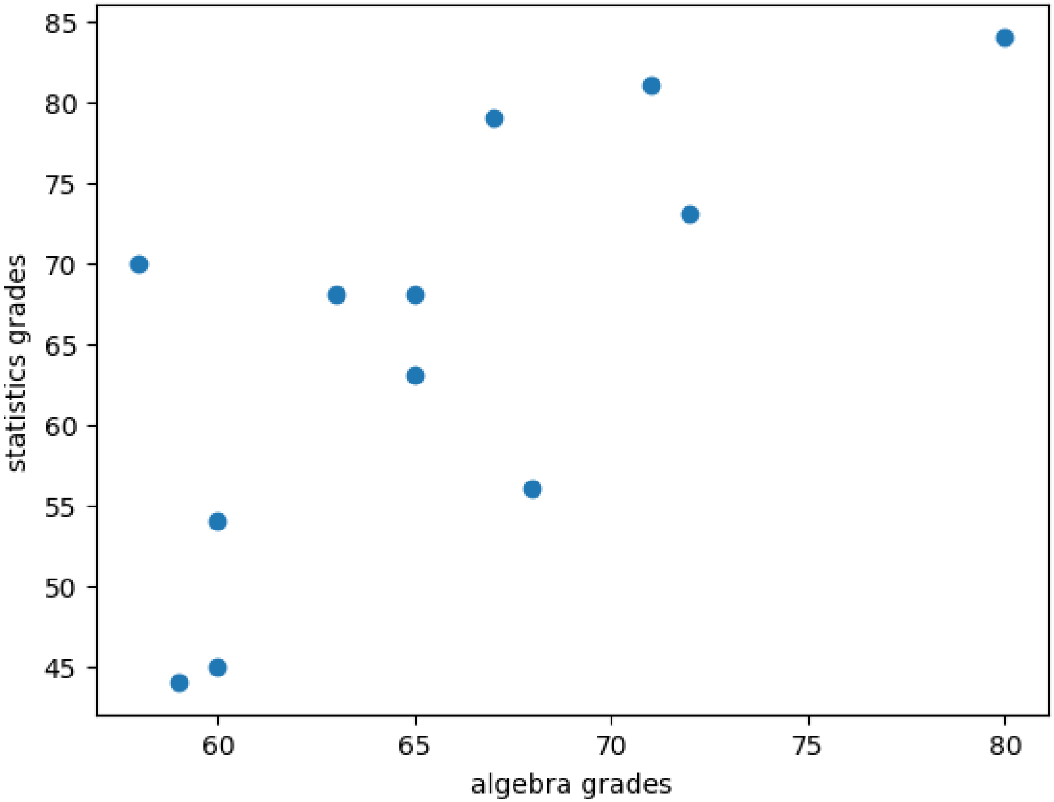
```
Algebra [67 80 71 63 65 72
65 68 58 60 60 59]
Statistics [79 84 81 68 63
73 68 56 70 45 54 44]
```



Example 1
Pearson correlation

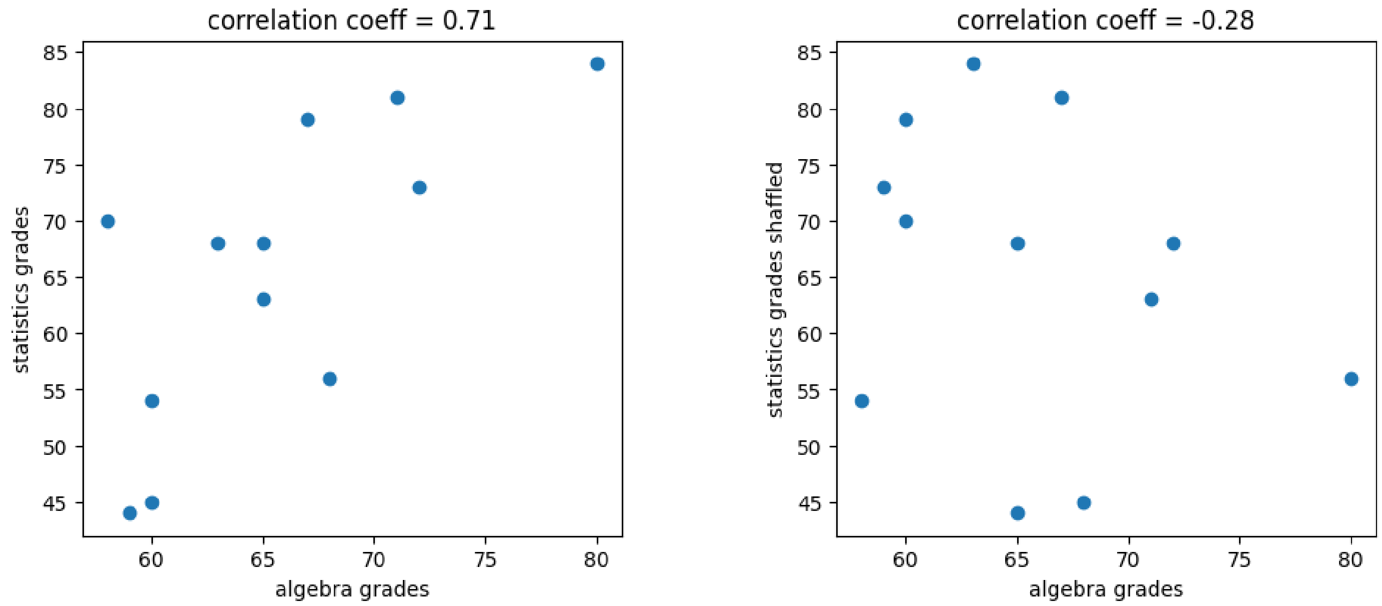
$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

```
Pearson correlation
coefficient: 0.71
```

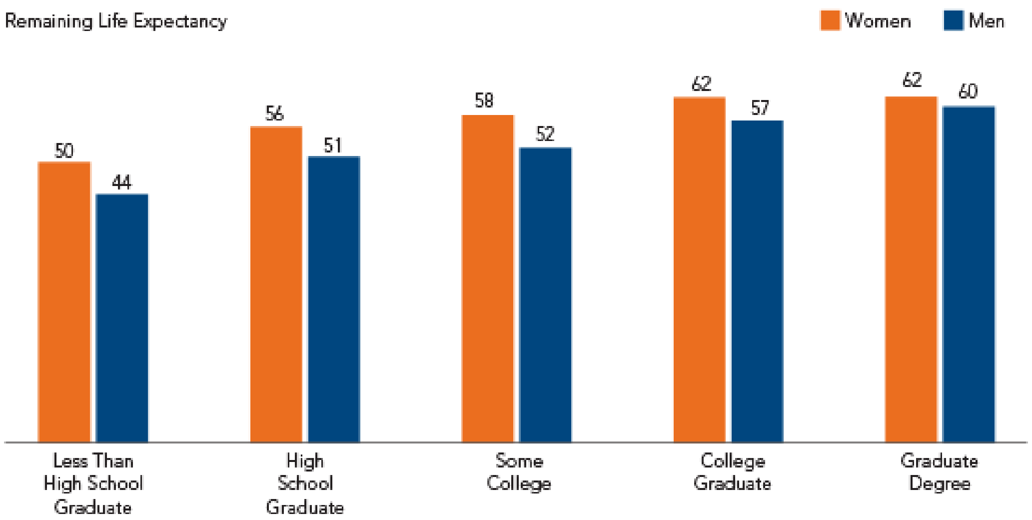


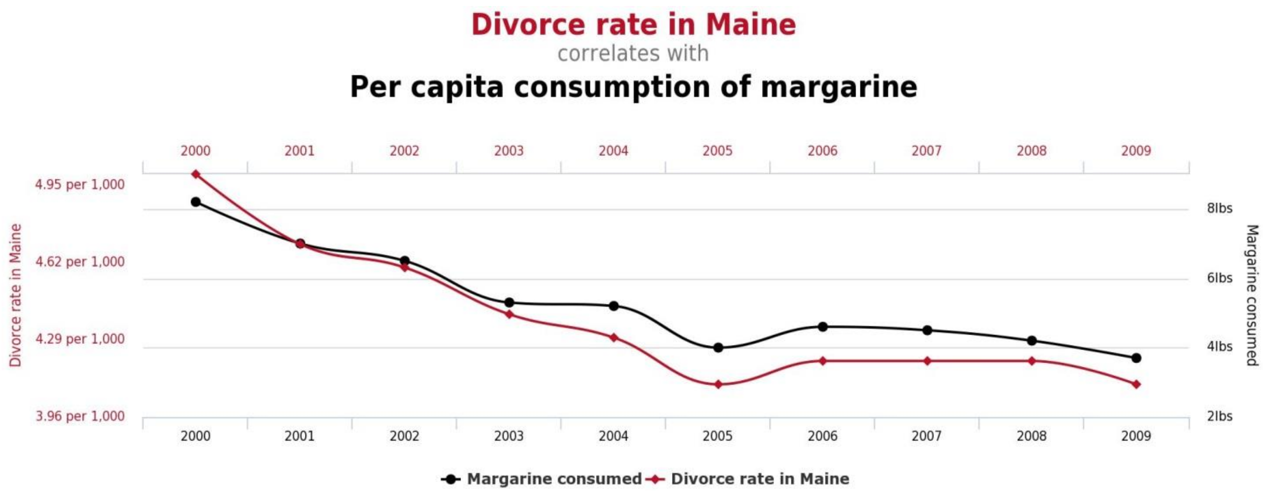
Example 1
correlated and uncorrelated data

We can shuffle one or both vectors to get uncorrelated vectors



Correlation ≠ causation





Linear regression

Data is represented with two variables

- Correlation is not enough

- We want to describe the relationship between the independent and dependent variables

Linear regression

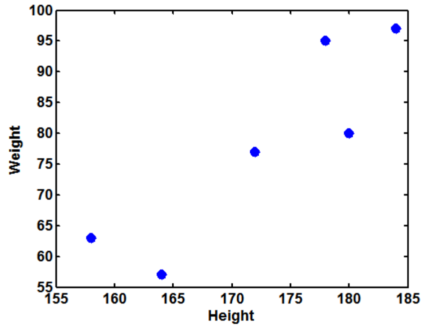
Population

$$y_n = \mu_{y|x_n} + \varepsilon_n$$

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

$$\varepsilon_n \sim N\left(0, \sigma^2_{y|x_n}\right)$$

Sample



Estimation

$$y_n = \hat{y}\left(x_n\right) + r_n$$

$$\hat{y}\left(x\right) = b_0 + b_1 x$$

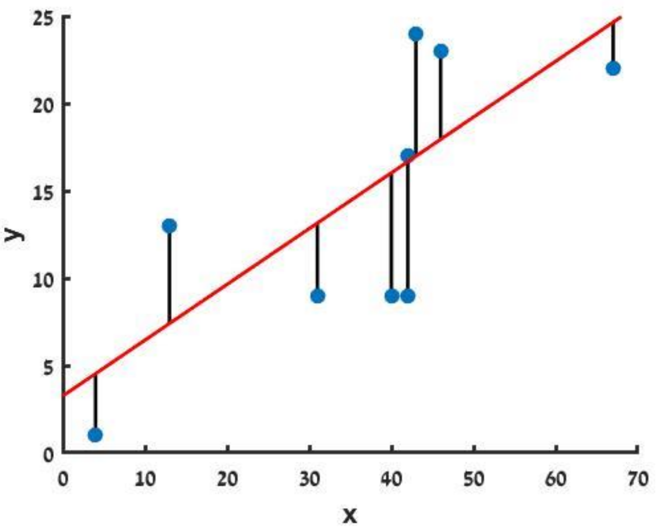
$$r_n \sim N\left(0, s^2_{y|x_n}\right)$$

b1 and b0 – regression coefficients

Estimating parameters

LSE – Least squares estimate

$$\hat{y}(x) = b_0 + b_1 x$$



$$SS_{\text{res}} = \sum_{n=1}^{N} \left(y_n - \hat{y}(x_n) \right)^2$$

$$= \sum_{n=1}^{N} \left(y_n - (b_0 + b_1 x_n) \right)^2$$

$$b_1 = \frac{\sum_{n=1}^{N} y_n x_n - N\overline{xy}}{\sum_{n=1}^{N} x_n{}^2 - N\overline{x}^2} = \frac{\sum_{n=1}^{N} (x_n - \overline{x})(y_n - \overline{y})}{\sum_{n=1}^{N} (x_n - \overline{x})^2} = \frac{\text{covariance}}{\text{variance}}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

Estimating variance

$$s_{y|x}^2 = \frac{SS_{res}}{N-2} = \frac{\sum_n (y_n - \hat{y}(x_n))^2}{N-2} = \frac{\sum_n (y_n - (b_0 + b_1 x))^2}{N-2}$$

$$se(b_1) = \frac{s_{y|x}}{\sqrt{\sum (x - \overline{x})^2}} = \frac{s_{y|x}}{\sqrt{(N-1)s_x{}^2}}$$

$$se_{b_0} = \sqrt{s_{\overline{y}}^2 + s_{b_1}^2 \overline{x}^2} = s_{y|x}\sqrt{\frac{1}{N} + \frac{\overline{x}^2}{\left(N-1\right)s_x^2}}$$

Variance estimator

- Model assumptions:
  - noise is normally distributed
  - variance is not dependent on X
- If assumptions are true, we use t-distribution
- If assumptions are not true, we use bootstrap

Exercise 2

We want to describe the relationship between age and blood pressure. We gathered data from 29 healthy people.

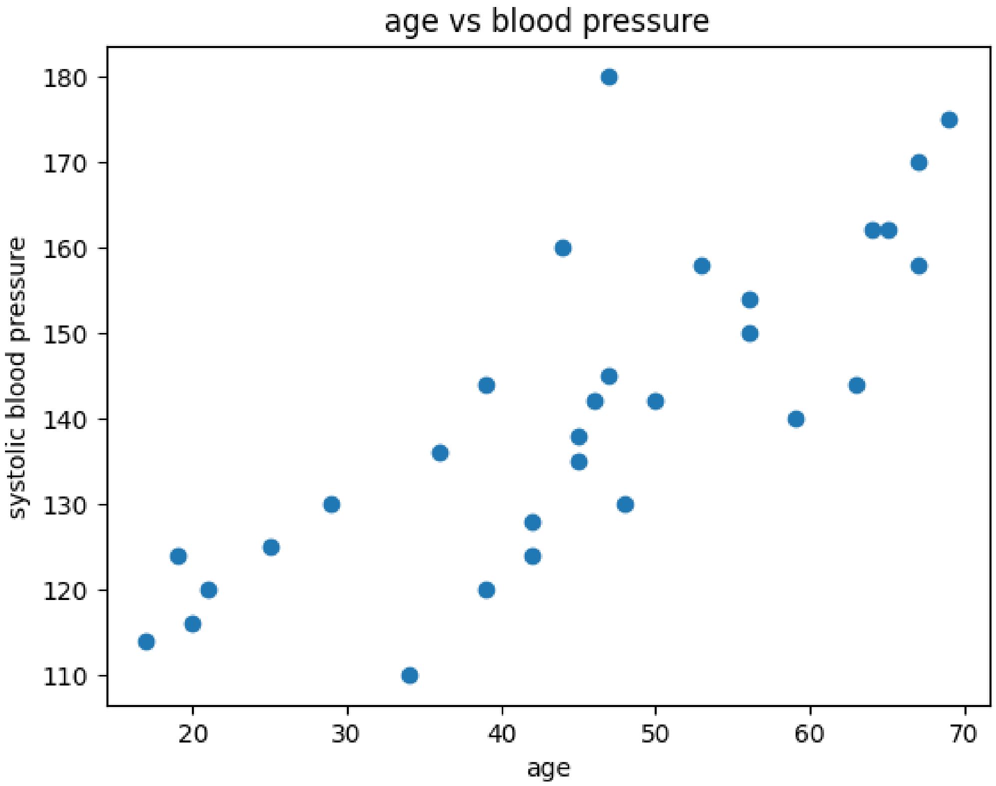
x – age

y – systolic blood pressure

subj	age	bp
1	39	144
2	47	180
3	45	138
4	47	145
5	65	162
6	46	142
7	67	170
8	42	124
9	67	158
10	56	154
11	64	162
12	56	150
13	59	140
14	34	110
15	42	128
16	48	130
17	45	135
18	17	114
19	20	116
20	19	124
21	36	136
22	50	142
23	39	120
24	21	120
25	44	160
26	53	158
27	63	144
28	29	130
29	25	125
30	69	175

Exercise 2

```
correlation coefficient: 0.79
```



$$\hat{y}(x) = b_0 + b_1 x$$

Linear Regression coefficient estimates

$$\hat{y}(x) = b_0 + b_1 x$$

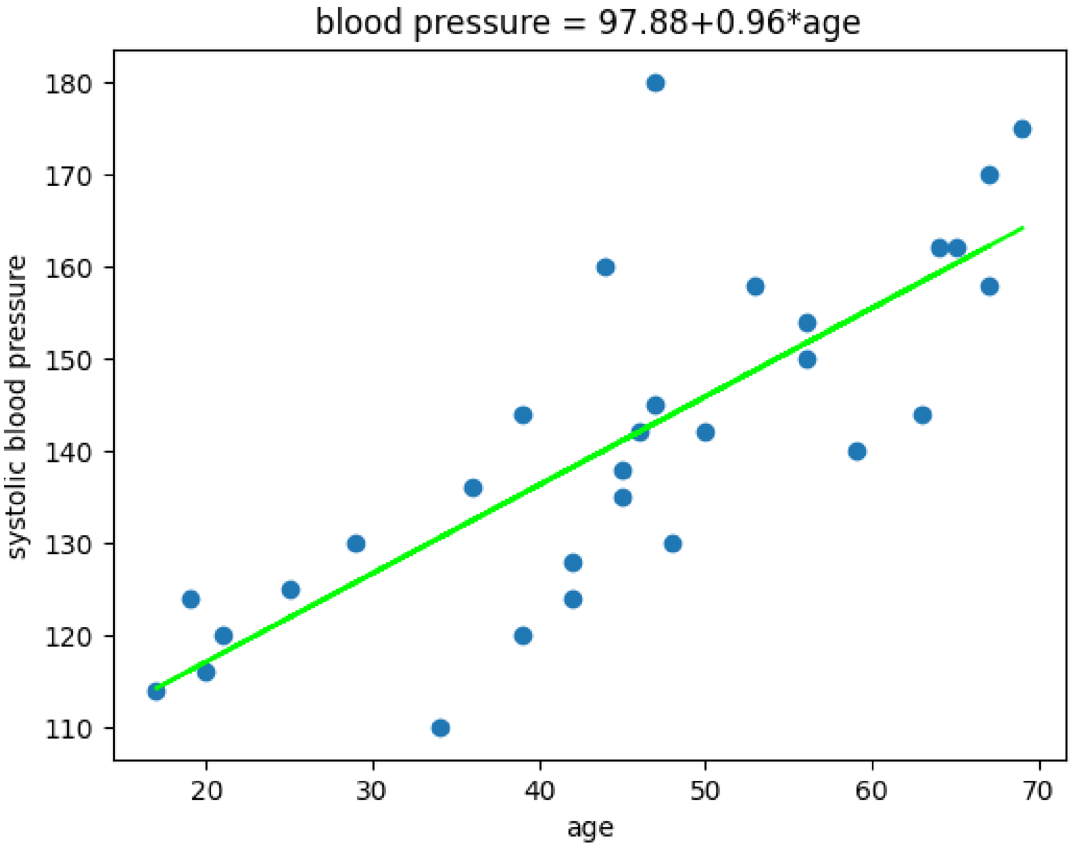
$$b_1 = \frac{\sum (x_n - \overline{x})(y_n - \overline{y})}{\sum (x_n - \overline{x})^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

b0 = 97.8782   b1 = 0.9599

Linear Regression Line

$$\hat{y}(x) = b_0 + b_1 x$$



Linear Regression Variance

$$se_{b_1} = \frac{s_{y|x}}{\sqrt{(N-1)s_x^2}}$$

$$s_{y|x}^2 = \frac{SS_{res}}{N-2} = \frac{\sum_n (y_n - \hat{y}(x_n))^2}{N-2}$$

$$se_{b_0} = s_{y|x}\sqrt{\frac{1}{N} + \frac{\overline{x}^2}{(N-1)s_x^2}}$$

Syx = 11.7822 Sb1 = 0.1431 Sb0 = 6.805

Linear Regression
b1 coefficient estimate CI

Calculate CI and significance using t-distribution

$$\frac{b_1 - \beta_1}{se_{b1}} \rightarrow t_{N-2}$$

$$b_1 - t_{N-2}{}^{\alpha/2} se_{b1} < \beta_1 < b_1 + t_{N-2}{}^{1-\alpha/2} se_{b1}$$

b1 95% CI = [0.6668 , 1.2529]

b1 p-value = 0.0

Linear Regression
b1 coefficient estimate CI

Calculate CI and significance using bootstrap:

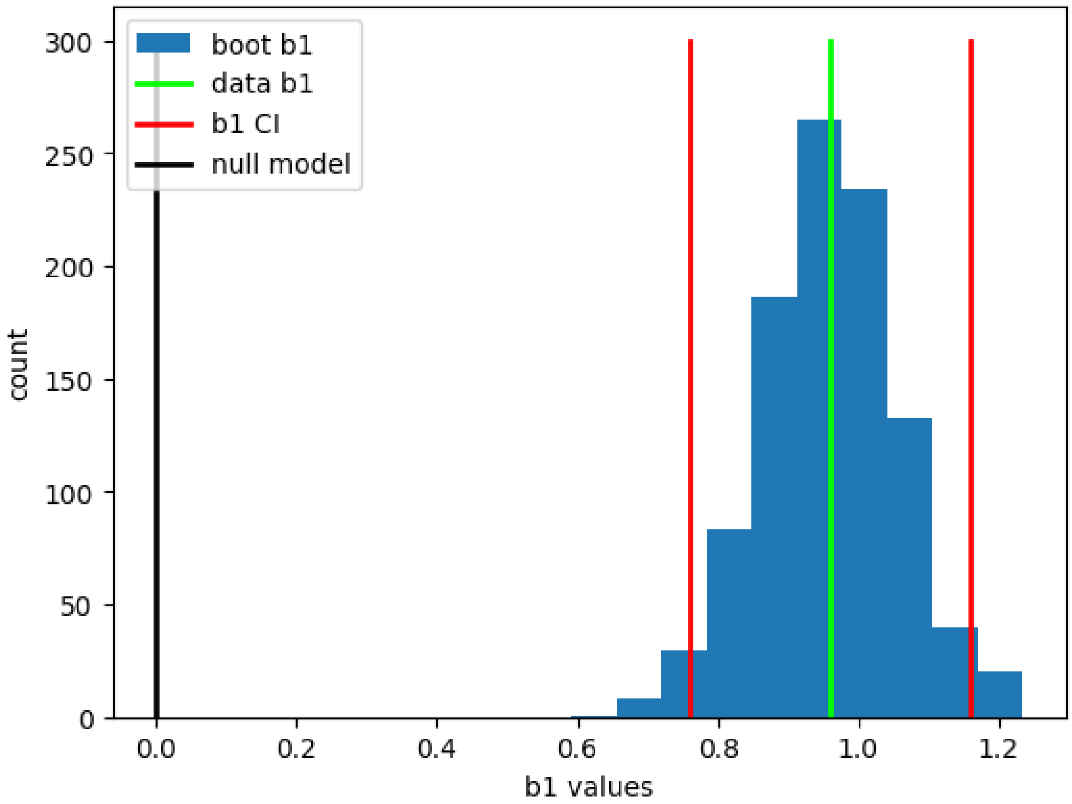
Create new sample of x and y vectors while keeping the connection

Calculate b1 coefficient for each pair of vectors

```
CI 95% for b1 coefficient = [0.7753 1.1527]

p-value for b1 coefficient = 0
```

Linear Regression
b1 coefficient estimate CI



Effect size

r^2 measure for a percentage of variance explained by the model

Range - between 0 and 1

f^2 Cohen's f : range – from 0 to infinity (f^2 <0.1 – small, f^2 >0.4 – large)

$$SS_{res} = \sum_i \left(y_i - \hat{y}_i \right)^2 \qquad SS_{reg} = \sum_i \left(\hat{y}_i - \overline{y}_i \right)^2 \qquad SS_{tot} = \sum_i \left(y_i - \overline{y}_i \right)^2$$

$$r^2 = \frac{SS_{\mathrm{reg}}}{SS_{\mathrm{tot}}} = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}} = b_1^2 \frac{s_x^2}{s_y^2} \qquad f^2 = \frac{SS_{\mathrm{reg}}}{SS_{\mathrm{res}}} = \frac{r^2}{1 - r^2}$$

```
R2 = 0.6166       f2 = 1.6079
```

Linear Regression Model Summary

$$y_n \sim b_0 + b_1 x_n + r_n \qquad y_n \sim 97.88 + 0.96 x_n + r_n$$

$$r_n \sim N(0, s^2_{y|x}) \qquad r_n \sim N(0{,}138.82)$$

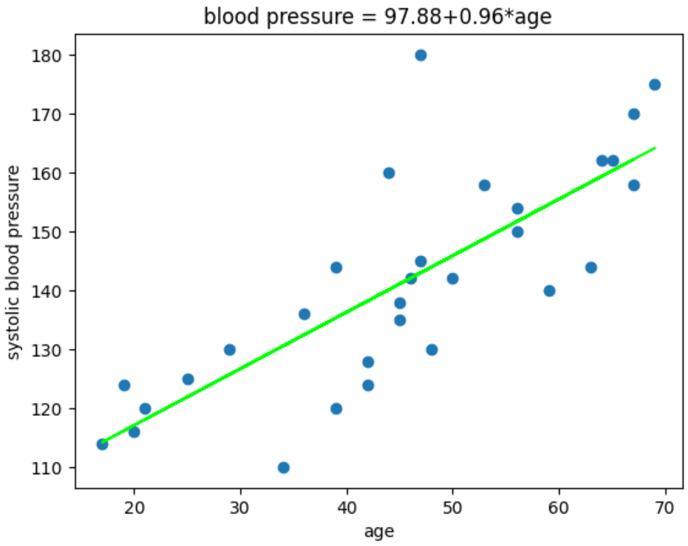
b0 = 97.8782

b1 = 0.9599 **CIb1 = 0.6668 1.2529** **p = 0**

Syx2 = 138.8192

R2 = 0.6166 **f2 = 1.6079**

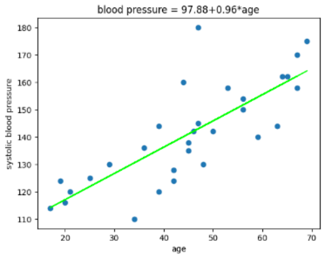
CI for Regression Line



$$\hat{y}(x) = b_0 + b_1 x$$

```
b1 95% CI = [0.6668 , 1.2529]
b0 95% CI = [83.9379 , 111.8184]
```

CI for Regression Line



- Calculate variance for regression line

$$\hat{y} = b_0 + xb_1 = \bar{y} - b_1\bar{x} + xb_1 = \bar{y} + (x - \bar{x})b_1$$

$$\mathrm{var}(\hat{y}) = \mathrm{var}(\bar{y} + (x - \bar{x})b_1) = \mathrm{var}(\bar{y}) + (x - \bar{x})^2 \,\mathrm{var}(b_1) = \frac{\mathrm{var}(y)}{N} + (x - \bar{x})^2 s_{b1}^2$$

$$se_{\hat{y}} = s_{y|x}\sqrt{\frac{1}{N} + \frac{(x - \bar{x})^2}{(N-1)s_x^2}}$$

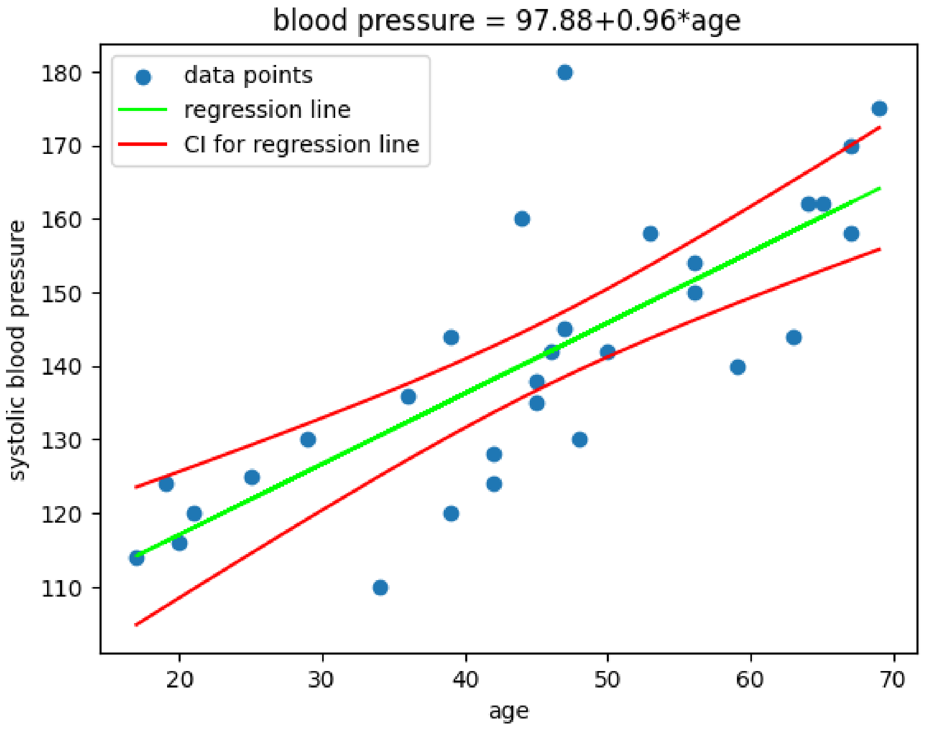
- Dependent on x
- We have more confidence around the middle and less – near the edges

$$\frac{\hat{y} - \mu_{y|x}}{se_{\hat{y}}} \to t_{N-2}; \qquad \hat{y} - t_{N-2}{}^{\alpha/2} se_{\hat{y}} < \mu_{y|x} < \hat{y} + t_{N-2}{}^{1-\alpha/2} se_{\hat{y}}$$

CI for Regression Line

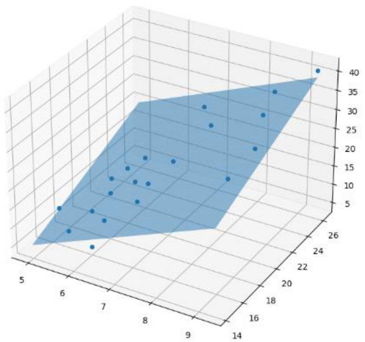
$$se_{\hat{y}} = s_{y|x}\sqrt{\frac{1}{N} + \frac{(x - \overline{x})^2}{(N-1)s_x^2}}$$

$$\hat{y} - t_{N-2}{}^{\alpha/2} se_{\hat{y}} < \mu_{y|x} < \hat{y} + t_{N-2}{}^{1-\alpha/2} se_{\hat{y}}$$



You can get the same results using libraries:
Scikit-learn
statsmodels

Multivariate Linear Regression



The model $$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon$$

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon \quad \text{where} \quad \mathbf{x} = \begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}; \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{Xb}$$

$$SS_{\mathrm{res}} = \sum \left(y_n - \hat{y}_n \right)^2 \qquad \hat{y}_n = b_0 + x_{n,1}b_1 + x_{n,2}b_2$$

$$\sum \left(y_n - \hat{y}_n \right)^2 = \left(\mathbf{y} - \mathbf{Xb} \right)^T \left(\mathbf{y} - \mathbf{Xb} \right)$$

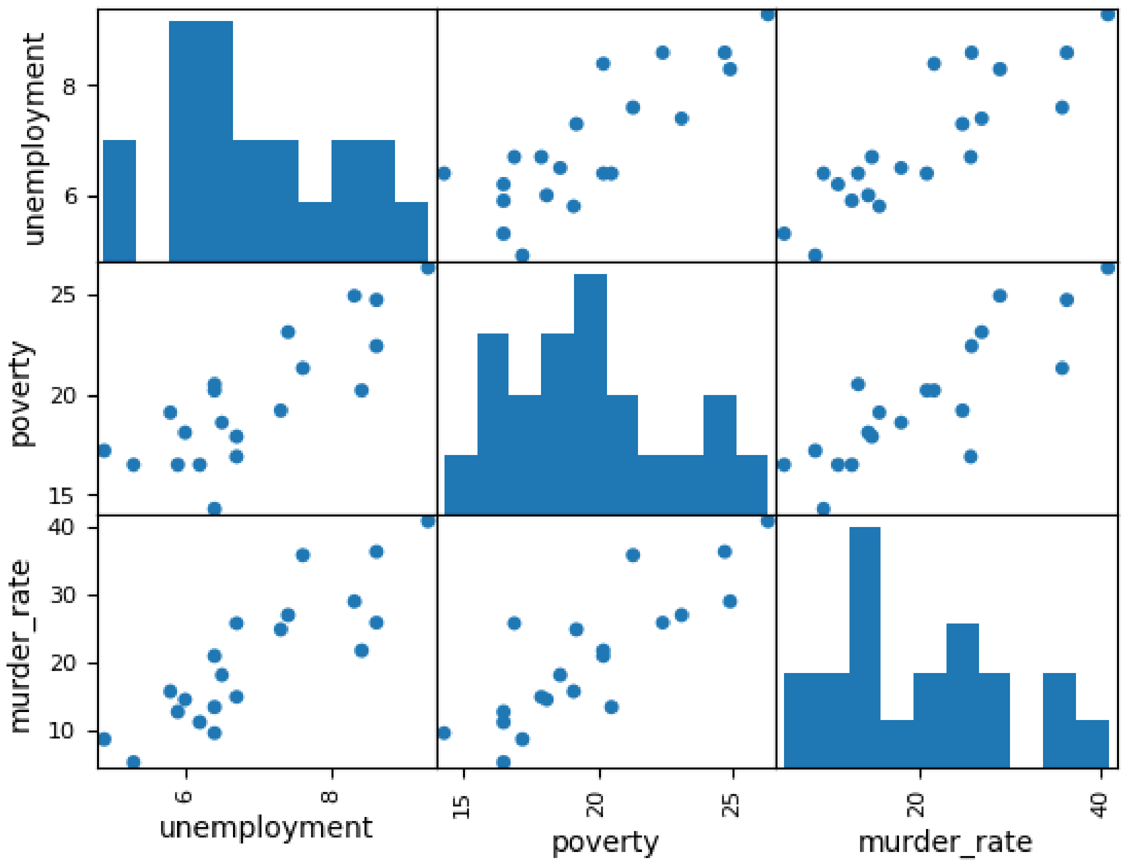
$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y} \qquad s_{y|x}^2 = MS_{\mathrm{res}} = \frac{SS_{\mathrm{res}}}{\nu_{\mathrm{res}}} = \frac{\left(\mathbf{y} - \mathbf{Xb} \right)^T \left(\mathbf{y} - \mathbf{Xb} \right)}{n - 3}$$

Example 3
Multivariate Linear Regression

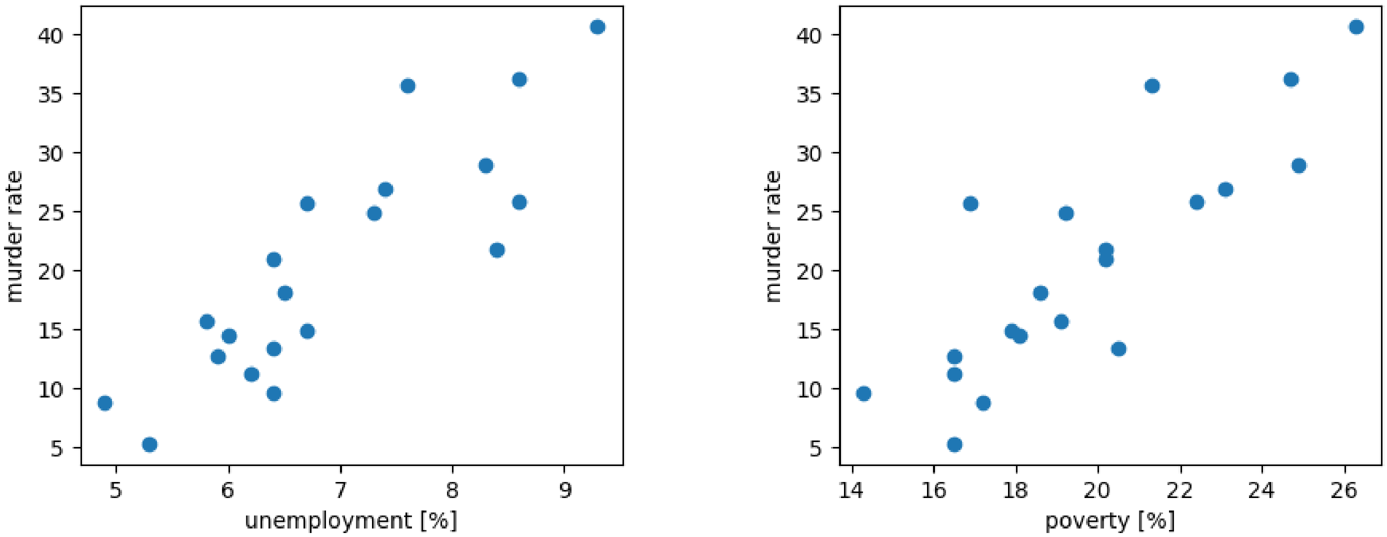
unemployment	poverty	murder_rate
6.2	16.5	11.2
6.4	20.5	13.4
9.3	26.3	40.7
5.3	16.5	5.3
7.3	19.2	24.8
5.9	16.5	12.7
6.4	20.2	20.9
7.6	21.3	35.7
4.9	17.2	8.7
6.4	14.3	9.6
6.0	18.1	14.5
7.4	23.1	26.9
5.8	19.1	15.7
8.6	24.7	36.2
6.5	18.6	18.1
8.3	24.9	28.9
6.7	17.9	14.9
8.6	22.4	25.8
8.4	20.2	21.7
6.7	16.9	25.7

In 1991, a group of researches gathered data on murder rate, poverty rate and unemployment in 20 different cities.
We will check the dependence of murder rate on poverty and unemployment

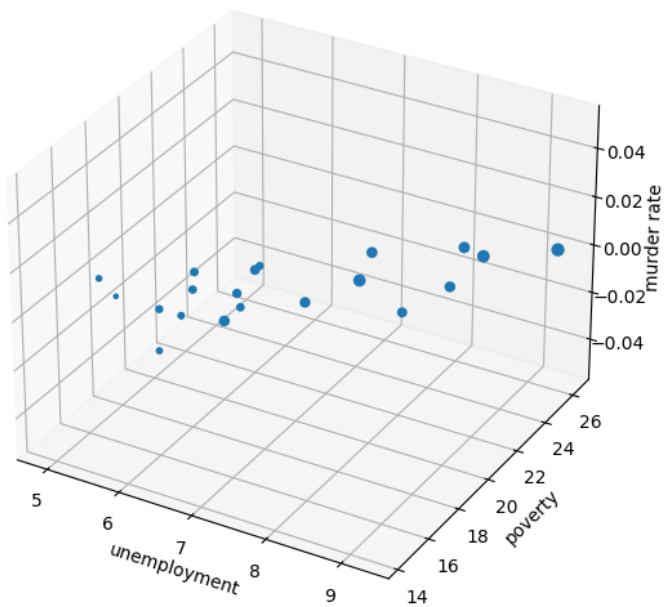
Example 3 presentation

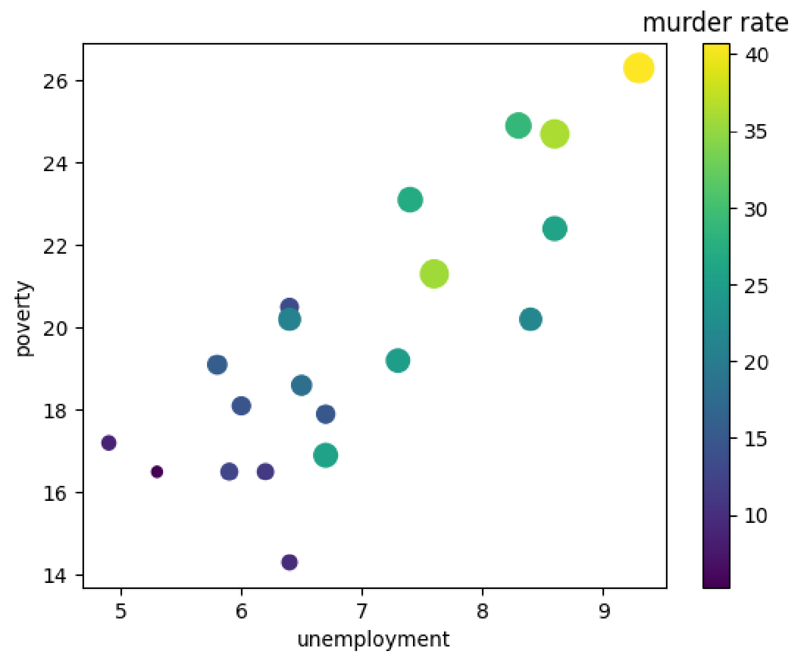


Example 3 presentation



Example 3 presentation



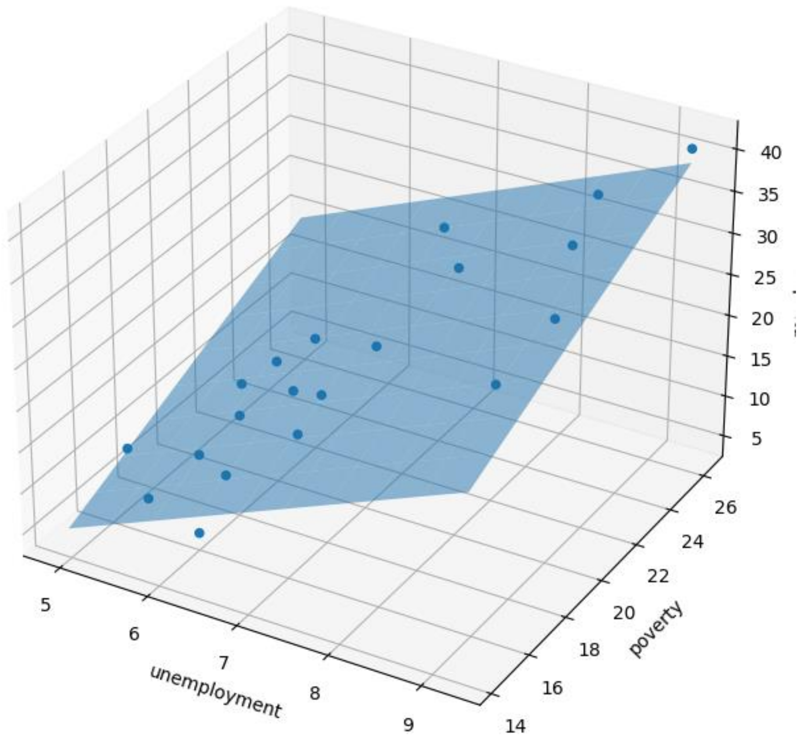


Example 3

$$\hat{y}_i = b_0 + b_1 x_i + b_2 x_i + r \quad ; \quad \mathrm{r} \sim \mathrm{N}(0, \mathrm{s}^2_{y|X})$$

b0 = -34.0725   b1 = 4.3989   b2 = 1.2239

Syx2 = 21.6084



Example 3

$$Cohen's\ \mathrm{f}^2 = \frac{r^2}{1-r^2}$$

effect size for regression

$$r^2 = 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}} = \frac{SS_{\mathrm{reg}}}{SS_{\mathrm{tot}}}$$

R2 = 0.802 f2 = 4.0503