

# Inferență statistică în ML

## Cap 6. Corelația și regresia liniară

April 22, 2023

- 1 Introducere în regresie
- 2 Metoda celor mai mici pătrate
- 3 Noțiunea de 'regresie către medie'

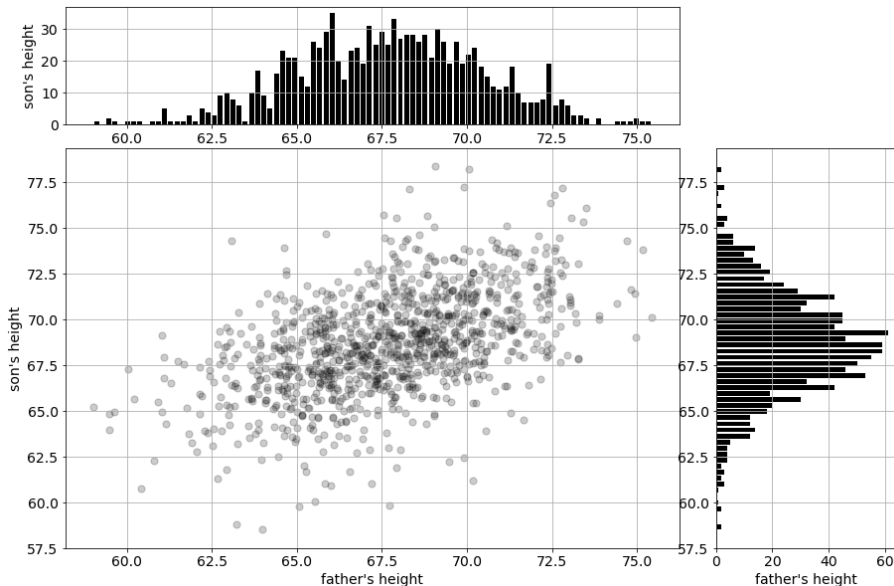
# Regression

- **regresia** este unul din modelele fundamentale, primul model de machine learning care se încearcă
- generalizarea sa este modelul liniar (Generalized Linear Model)
- concept (împreună cu modelul de corelație liniară) inventat de matematicianul britanic Francis Galton (sec XIX), pentru predicția a înălțimii fiilor folosind înălțimea taților (father.son dataset)
- explică relația simplă de medie dintre înălțimile fiilor și înălțimile taților într-un mod universal valabil (model transparent vs. NN = black-box)
- investighează variația înălțimii fiilor care pare să nu fie dependentă de înălțimea taților (variația reziduală)
- cuantifică ce impact are înălțimea taților în explicarea înălțimii fiilor
- dă un set de presupuneri care trebuie să fie valabile pentru ca modelul să generalizeze bine în afara sample-ului
- principiul 'regresiei către medie' ( $T \rightarrow t, S \leftarrow s$ )

# Galton dataset: father.son

- dataset folosit în 1885 de Francis Galton (1822 - 1911)
- Galton a inventat conceptele de regresie, corelație, deviație standard, percentile
- vărul lui Charles Darwin
- <https://select-statistics.co.uk/blog/regression-to-the-mean-as-relevant-today-as-it-was-in-the>
- vom vedea distribuțiile marginale ale taților -  $P(\text{father})$ , respectiv ale fiilor -  $P(\text{son})$ , respectiv cea bidimensională  $P(\text{father}, \text{son})$

# Distribuțiile marginale și centrală



# Găsirea mijlocului prin metoda least squares

- metoda celor mai mici pătrate (least squares)
- vom considera doar distribuția marginală a fiilor
- dacă  $Y_i$  sunt valorile înălțimilor fiilor pentru  $i = 1 \dots n$ , 'mijlocul' este dat de valoarea  $\mu$  care minimizează:

$$\sum_{i=1}^n (Y_i - \mu)^2$$

- mijlocul este centrul de masă al histogramei, respectiv valoarea medie  $\mu = \bar{Y}$

## Găsirea mijlocului prin metoda least squares (2)

$$S(\mu) = \sum_{i=1}^n (Y_i - \mu)^2$$

$$\min_{\mu} S(\mu) = \min_{\mu} \sum_{i=1}^n (Y_i - \mu)^2$$

$$\frac{\partial S}{\partial \mu} = -2 \sum_{i=1}^n (Y_i - \mu) = 0$$

$$\sum_{i=1}^n Y_i - n\mu = 0 \quad \text{adică} \quad \mu = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

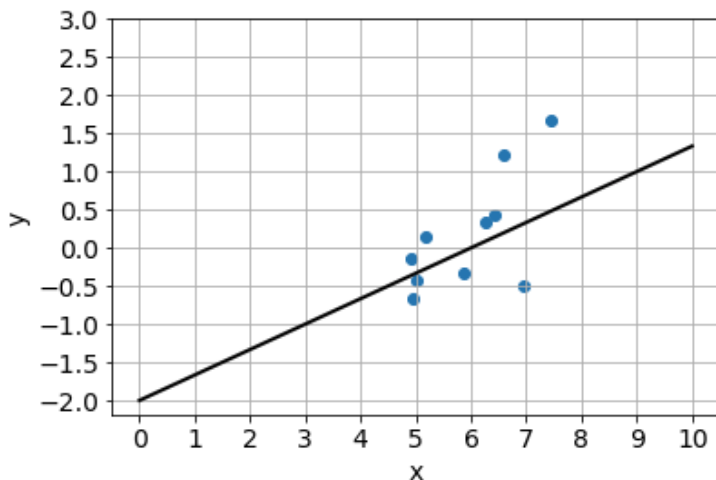
## Găsirea mijlocului prin metoda least squares (variantă)

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \mu)^2 &= \sum_{i=1}^n (Y_i - \bar{Y} + \bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n [(Y_i - \bar{Y})^2 + 2(Y_i - \bar{Y})(\bar{Y} - \mu) + (\bar{Y} - \mu)^2] \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \sum_{i=1}^n (Y_i - \bar{Y}) + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + 2(\bar{Y} - \mu) \underbrace{\left( \sum_{i=1}^n Y_i - n\bar{Y} \right)}_{=0} + \sum_{i=1}^n (\bar{Y} - \mu)^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n (\bar{Y} - \mu)^2 \geq \sum_{i=1}^n (Y_i - \bar{Y})^2 \Rightarrow \mu = \bar{Y}
\end{aligned}$$



- 1 Introducere în regresie
- 2 Metoda celor mai mici pătrate**
- 3 Noțiunea de 'regresie către medie'

# Problema



- determinarea expresiei funcției  $f(x) \sim y$

# Notatii pentru date

- cele  $n$  puncte sunt descrise ca  $X_1, X_2 \dots X_n$
- de exemplu, pentru setul  $X = \{4, 6, 3\}$  avem  $X_1 = 4, X_2 = 6, X_3 = 3$
- putem folosi pentru puncte notații precum  $X_i, Y_j$  etc.
- cantitățile pe care nu le cunoaștem, dar dorim să le estimăm, sunt notate cu litere grecești, de exemplu  $\mu, \beta \dots$
- ne reamintim că o variabilă aleatoare  $X \sim N(\mu, \sigma)$  se poate normaliza (standardiza sau centra):  $\frac{X-\mu}{\sigma} \sim N(0, 1)$

# Deviația standard empirică

- dispersia empirică este definită ca:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

- deviația standard empirică este  $S = \sqrt{S^2}$ ; deviația standard are aceeași unitate de măsură ca a datelor
- normalizarea datelor (centrare + scalare):  $Z_i = \frac{X_i - \bar{X}}{S} \sim N(0, 1)$

# Covariance

- **covariația** empirică este caracteristică perechilor de variabile aleatoare care iau același număr de valori și este descrisă de:

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} \right) \end{aligned}$$

- covariația descrie cum variază abaterea față de media proprie sincron
- **corelația** descrie același lucru dar scalat la produsul dispersiilor variabilelor aleatoare:

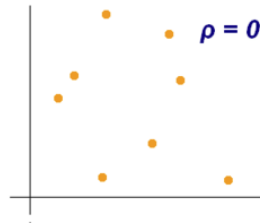
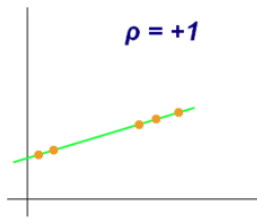
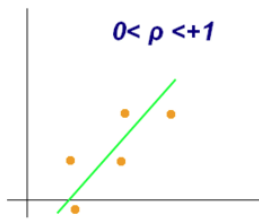
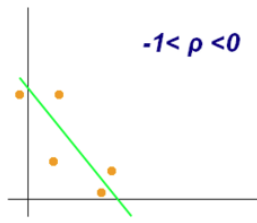
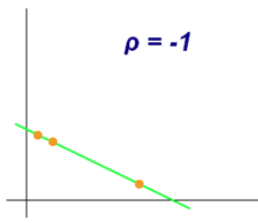
$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{S_x S_y}$$

- din cauza scalării, corelația are valoarea cuprinsă în intervalul  $[-1, 1]$

# Proprietățile corelației

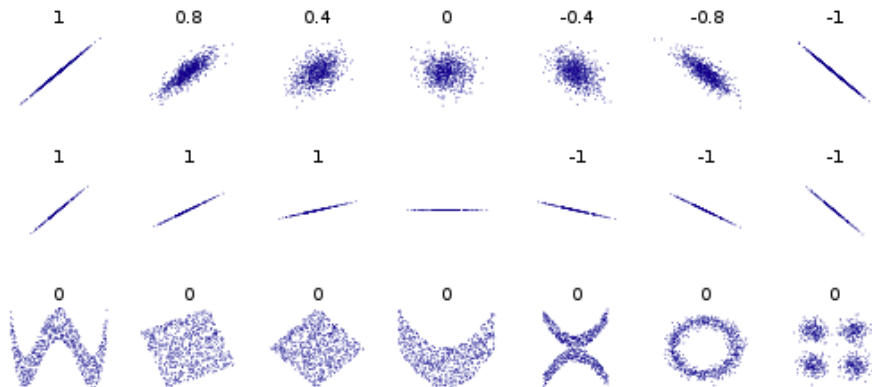
- $Cor(X, Y) = Cor(Y, X)$
- $-1 \leq Cor(X, Y) \leq 1$
- $Cor(X, Y) = 1$  și  $Cor(X, Y) = -1$  doar când observațiile descrise  $X$  și  $Y$  descriu o dreaptă cu pantă pozitivă, respectiv negativă
- $Cor(X, Y)$  măsoară tăria dependenței liniare dintre  $X$  și  $Y$ ; cu cât mai puternică dependența liniară, cu atât corelația se apropie de  $-1$  sau  $1$
- $Cor(X, Y) = 0$  implică inexistența unei corelații liniare între  $X$  și  $Y$

# Coeficientul de corelație<sup>1</sup> (Pearson)



<sup>1</sup>[https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)

# Coeficientul de corelație<sup>2</sup> nu reflectă nici panta nici forma



<sup>2</sup>[https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)



# General least squares

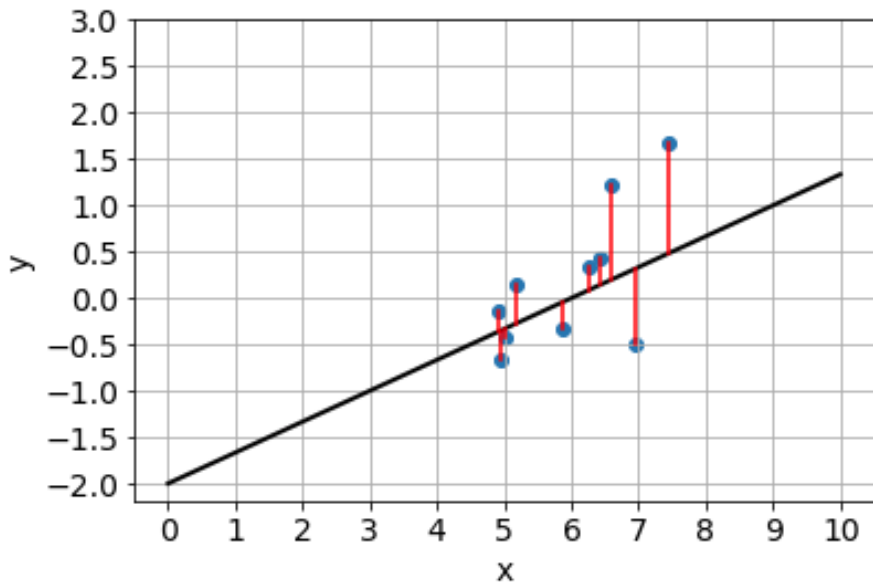
- dorim să exprimăm înălțimea copiilor în funcție de înălțimea părinților
- construim așadar predicția  $y \sim x$ , folosind regresia liniară
- vom construi 'cea mai potrivită' linie
- considerăm  $Y_i$  înălțimile copiilor și  $X_i$  înălțimile părinților
- cea mai potrivită linie presupune minimizarea erorii survenite (metoda celor mai mici pătrate):

$$\sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

$$f(X_i) = \beta_0 + \beta_1 X_i$$

- parametrii regresiei sunt  $\beta_1$ , denumit slope, și  $\beta_0$ , denumit intercept

# Principiul least squares



## Principiul least squares (2)

- vom căuta soluțiile  $\beta_1, \beta_0$  ce minimizează expresia:

$$\min_{\beta_1, \beta_0} J(\beta_1, \beta_0) = \min_{\beta_1, \beta_0} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \quad (1)$$

- cum este vorba de o problemă de optim, acesta se află prin anularea derivatelor parțiale:

$$0 = \frac{\partial J}{\partial \beta_0} = 2 \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)] \quad \text{respectiv}$$

$$0 = \frac{\partial J}{\partial \beta_1} = 2 \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)] (-X_i)$$

# Principiul least squares (3)

- din prima relație obținem:

$$\begin{aligned}
 0 &= \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)] \\
 &= \sum_{i=1}^n Y_i - n\beta_0 - \beta_1 \sum_{i=1}^n X_i \\
 &= n\bar{Y} - n\beta_0 - n\beta_1\bar{X} \\
 &= n(\bar{Y} - \beta_0 - \beta_1\bar{X})
 \end{aligned}$$

- adică:

$$\boxed{\beta_0 = \bar{Y} - \beta_1\bar{X}} \quad (2)$$

## Principiul least squares (4)

- din a doua relație obținem (am redus cu -2):

$$\begin{aligned}
 0 &= \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)] X_i \\
 &= \sum_{i=1}^n X_i Y_i - \beta_0 \sum_{i=1}^n X_i - \beta_1 \sum_{i=1}^n X_i^2 \\
 &= \sum_{i=1}^n X_i Y_i - n\beta_0 \bar{X} - \beta_1 \sum_{i=1}^n X_i^2 \quad , \text{introducem relația pentru } \beta_0: \\
 &= \sum_{i=1}^n X_i Y_i - n(\bar{Y} - \beta_1 \bar{X}) \bar{X} - \beta_1 \sum_{i=1}^n X_i^2 \\
 &= \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} + n\beta_1 \bar{X}^2 - \beta_1 \sum_{i=1}^n X_i^2
 \end{aligned}$$

## Principiul least squares (5)

$$\begin{aligned}
&= \sum_{i=1,n} X_i Y_i - n\bar{X}\bar{Y} + n\beta_1\bar{X}^2 - \beta_1 \sum_{i=1,n} X_i^2 \\
&= \sum_{i=1,n} X_i Y_i - n\bar{X}\bar{Y} - n\beta_1 \left( \frac{1}{n} \sum_{i=1,n} X_i^2 - \bar{X}^2 \right) \\
&= \sum_{i=1,n} X_i Y_i - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} - n\bar{X}\bar{Y} - n\beta_1 \text{Var}(X) \\
&= \sum_{i=1,n} X_i Y_i - \sum_{i=1,n} \bar{X} Y_i + \sum_{i=1,n} \bar{X}\bar{Y} - \sum_{i=1,n} X_i \bar{Y} - n\beta_1 \text{Var}(X) \\
&= \sum_{i=1,n} (X_i - \bar{X}) Y_i - \sum_{i=1,n} (X_i - \bar{X}) \bar{Y} - n\beta_1 \text{Var}(X) \\
&= \sum_{i=1,n} (X_i - \bar{X})(Y_i - \bar{Y}) - n\beta_1 \text{Var}(X) = n\text{Cov}(X, Y) - n\beta_1 \text{Var}(X)
\end{aligned}$$

# Principiul least squares (6)

- adică:

$$nCov(X, Y) - n\beta_1 Var(X) = 0$$

$$\beta_1 Var(X) = Cov(X, Y)$$

$$\beta_1 Std(X)Std(X) = Cor(X, Y)Std(X)Std(Y)$$

- sau:

$$\beta_1 = Cor(X, Y) \frac{Std(Y)}{Std(X)}$$

# Least squares - sumar

- modelul realizează potrivirea (fit) funcției liniare  $Y \sim \hat{\beta}_0 + \hat{\beta}_1 X$ , unde:

$$\hat{\beta}_1 = \text{Cor}(X, Y) \frac{\text{Std}(Y)}{\text{Std}(X)} \quad \text{și} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- coeficienții  $\beta_0$  și  $\beta_1$  nu sunt cei exacti, ci doar niște estimatori calculați pe sample-ul nostru
- $\hat{\beta}_1$  este în unități de  $Y/X$ , iar  $\hat{\beta}_0$  are unitatea de măsură a lui  $Y$
- deoarece  $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$ , dreapta de regresie trece prin punctul  $(\bar{X}, \bar{Y})$
- panta regresiei este aceeași cu cea obținută dacă am standardiza centra datele,  $(X_i - \bar{X}, Y_i - \bar{Y})$ , dreapta regresiei trece prin origine
- dacă am și standardiza datele, panta regresiei devine  $\text{Cor}(X, Y)$



# Codul pentru regresia liniară

```
# see https://machinelearningmastery.com/introduction-to-expected-value-variance-and-covariance/

x, y = np.array(fheight), np.array(sheight)
beta1 = np.corrcoef(x, y)[0, 1] * np.std(y)/np.std(x)
beta0 = np.mean(y) - beta1 * np.mean(x)
print('beta0:', beta0, 'beta1:', beta1)

xext = sm.add_constant(x)

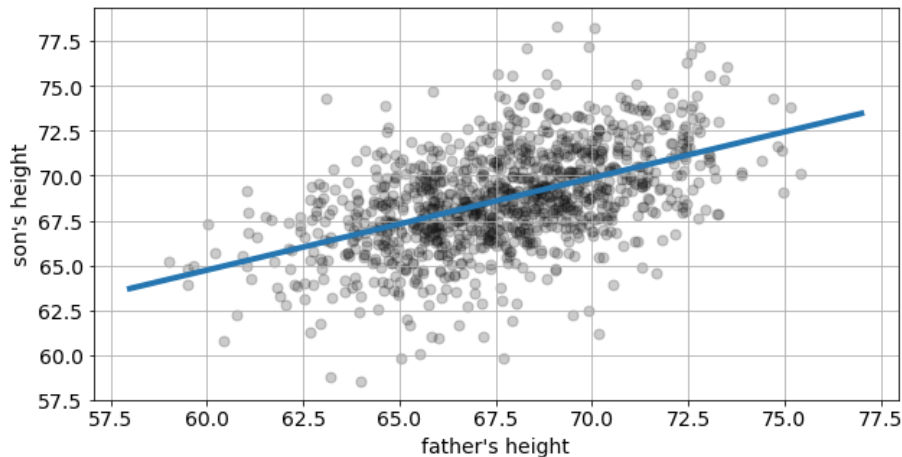
lm = sm.OLS(y, xext).fit()
print('intercept:', lm.params[0], 'coefficient:', lm.params[1])

x1 = np.linspace(58, 77, 100)
y1 = beta1 * x1 + beta0

fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax.scatter(x, y, c='k', alpha = .2, s=50)
ax.plot(x1, y1, lw=4)
ax.set(xlabel="father's height", ylabel="son's height")
ax.grid(True)
plt.show()
```

```
beta0: 33.88660435407794 beta1: 0.5140930386233075
intercept: 33.88660435407793 coefficient: 0.514093038623308
```

# Dreapta de regresie



- 1 Introducere în regresie
- 2 Metoda celor mai mici pătrate
- 3 Noțiunea de 'regresie către medie'

# Regresia către medie

- idee faimoasă avansată de Francis Galton
- de ce copiii părinților înalți tind să fie înalți, dar nu la fel de înalți ca și părinții lor?
- de ce copiii părinților scunzi sunt și ei scunzi, dar totuși mai înalți decât părinții lor?
- de ce părinții copiilor înalți sunt și ei înalți, dar nu așa înalți ca și copiii lor?
- de ce Simona Halep, care a ajuns top1 mondial, tinde să joace mai prost anul ăsta?
- de ce prețurile unor acțiuni, care au mers așa bine, acum scad?
- este un motiv intrinsec sau este doar regresia către mediocritate?

## Regresia către medie (2)

- toate sunt exemple de fenomene de tip 'regresie către medie'

```
x = np.random.randn(10)
ordered = np.argsort(x)[::-1]

y = np.random.randn(10)

# doar in 1 caz din 11 vom obtine un numar mai mic
print(x[ordered[0]], '>', y[ordered[0]])

1.6560722892507018 > -2.3249257309267626
```

```
# verificare
n = 10000000
x = np.random.randn(n, 10)
ordered = np.argsort(x, axis=1)
for i in range(x.shape[0]):
    x[i, :] = x[i, ordered[i][::-1]]

y = np.random.randn(n, 10)
print(np.sum(x[:, 0] < y[:, 0]) / n, 'vs. 1/11 =', 1/11)

0.0909046 vs. 1/11 = 0.09090909090909091
```

- maximul dintr-un set de 10 numere are o șansă de 10 din 11 să fie mai mare decât numărul de pe aceeași poziție din al doilea set

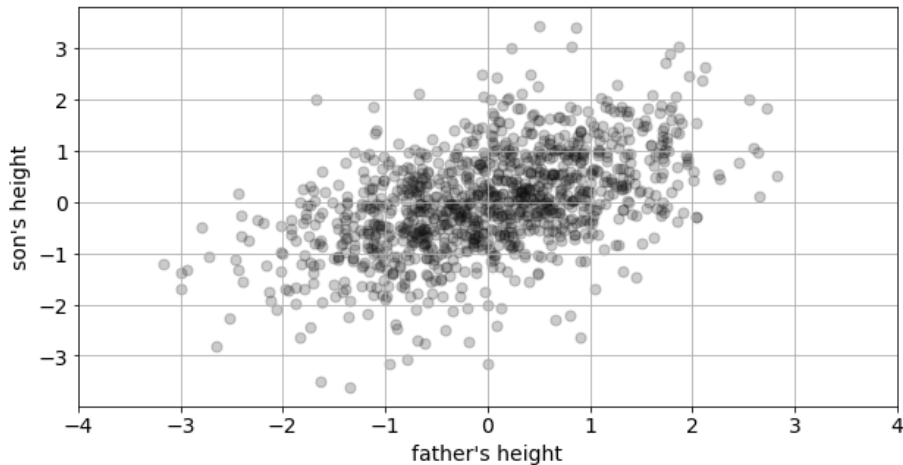
## Regresia către medie (3)

- pentru două seturi de numere extrase aleator, cel mai mare număr din primul set va fi mai mare din pură întâmplare
- probabilitatea ca în al doilea set să fie altele mai mici decât acest maxim este mare
- adică  $P(Y < X | X = x)$  crește atunci când  $x$  merge spre valori mari
- similar,  $P(Y > X | X = x)$  crește pe măsură ce  $x$  merge spre valori mai mici
- ne putem gândi la dreapta de regresie ca la un instrument care indică dependența liniară de valorile anterioare
- în cazul  $Cor(Y, X) = 1$ , atunci  $Y$  se comportă la fel ca  $X$ , altfel, substratul nu e identic

## Regresia către medie (4)

- presupunem că normalizăm atât  $X$  (înălțimea părinților) precum și înălțimea copiilor ( $Y$ ), astfel ca ambele să aibă media 0 și deviația standard 1
- în aceste condiții dreapta de regresie trece prin origine ( $\bar{X}, \bar{Y}$ )
- panta dreptei de regresie este atunci  $Cor(Y, X)$

# Regresia către medie (5)





# Regresia către medie (6)

```
# originea este acum (Xbar, Ybar)
x, y = np.array(fheight), np.array(sheight)
x, y = (x - np.mean(x))/np.std(x), (y - np.mean(y))/np.std(y)
rho = np.corrcoef(x, y)[0, 1]
print('rho:', rho)

x1 = np.linspace(-4, 4, 100)

fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax.scatter(x, y, c='k', alpha = .2, s=50)
ax.plot(x1, rho * x1, lw=4)
ax.plot([2, 2, -4], [0, 2, 2], 'g', lw=2)
ax.plot([2, 2, -4], [0, 1, 1], 'r', lw=2)
ax.plot(x1, x1, lw=1, c='k')
ax.plot(x1, [0] * len(x1), lw=1, c='k')
ax.legend(['dreapta de regresie', 'fără zgomet,\nncorelație perfectă', 'regresia către medie',
          'doar zgomet,\n nicio corelație'])
ax.set(xlabel="father's height", ylabel="son's height", xlim=(-4, 4))
ax.grid(True)
plt.show()
```

rho: 0.5013383111723431

# Regresia către medie (7)

