

# Vorlesung 8a

Kovarianz und Korrelation,

Regressionsgerade

# 1. Die Kovarianz und ihre Eigenschaften

Wir erinnern an die Definition der  
**Kovarianz**

Für reellwertige Zufallsvariable  $X, Y$   
mit  $\mathbf{E}[X^2] < \infty$  und  $\mathbf{E}[Y^2] < \infty$  ist

$$\mathbf{Cov}[X, Y] := \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]$$

Insbesondere ist also

$$\mathbf{Cov}[X, X] = \mathbf{Var}[X]$$

Die Kovarianz ist

- im Fall von zwei gleichen Einträgen nichtnegativ:

$$\text{Cov}[X, X] \geq 0$$

- in den beiden Einträgen symmetrisch:

$$\text{Cov}[X, Y] = \text{Cov}[Y, X]$$

- *bilinear*, d.h. in jedem einzelnen Eintrag linear:

$$\text{Cov}[c_1X_1 + c_2X_2, Y] = c_1\text{Cov}[X_1, Y] + c_2\text{Cov}[X_2, Y]$$

## 2. Die Kovarianz-Varianz-Ungleichung

Die “Kovarianz-Varianz-Ungleichung”

$$|\text{Cov}[X, Y]| \leq \sqrt{\text{Var}X} \sqrt{\text{Var}Y}$$

folgt (mit  $G := X - \mu_X$ ,  $H := Y - \mu_Y$ ) sofort aus der

**Cauchy-Schwarz Ungleichung:**

Für reellwertige Zufallsvariable  $G, H$  mit  $\mathbf{E}[G^2], \mathbf{E}[H^2] < \infty$

ist

$$(\mathbf{E}[GH])^2 \leq \mathbf{E}[G^2] \mathbf{E}[H^2] .$$

Behauptung:  $\pm \mathbf{E}[GH] \leq \sqrt{\mathbf{E}[G^2]} \sqrt{\mathbf{E}[H^2]}$

Beweis:

Fall 1:  $\mathbf{E}[G^2], \mathbf{E}[H^2] > 0$ .

$U := G/\sqrt{\mathbf{E}[G^2]}, V := H/\sqrt{\mathbf{E}[H^2]}$  erfüllen

$$\mathbf{E}[U^2] = \mathbf{E}[V^2] = 1.$$

Aus  $\pm 2UV \leq U^2 + V^2$  folgt  
 $\pm \mathbf{E}[UV] \leq \frac{1}{2}(\mathbf{E}[U^2] + \mathbf{E}[V^2]) = 1.$

$$\pm \frac{\mathbf{E}[GH]}{\sqrt{\mathbf{E}[G^2]} \sqrt{\mathbf{E}[H^2]}} \leq 1.$$

Behauptung:  $\pm \mathbf{E}[GH] \leq \sqrt{\mathbf{E}[G^2]} \sqrt{\mathbf{E}[H^2]}$

Fall 2:  $\mathbf{E}[G^2] = 0$ .

Dann folgt aus dem

Satz von der Positivität des Erwartungswertes

$$\mathbf{P}(G^2 = 0) = 1,$$

$$\text{also } \mathbf{P}(GH = 0) = 1$$

und

$$\mathbf{E}[GH] = 0. \quad \square$$

### 3. Der Korrelationskoeffizient

## Definition.

Für zwei Zufallsvariable  $X, Y$   
mit positiven, endlichen Varianzen ist

$$\kappa_{XY} := \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}X} \sqrt{\text{Var}Y}}$$

der **Korrelationskoeffizient** von  $X$  und  $Y$ .

Aus der Kovarianz-Varianz-Ungleichung folgt sofort

$$-1 \leq \kappa_{XY} \leq 1.$$

## 4. Die Bedeutung des Korrelationskoeffizienten

Fünf prominente Zahlen  
zur (teilweisen) Beschreibung der Verteilung  
eines zufälligen Paares  $(X, Y)$  in  $\mathbb{R} \times \mathbb{R}$ :

$\mu_X$  und  $\mu_Y$ : die Erwartungswerte von  $X$  und  $Y$

$\sigma_X$  und  $\sigma_Y$ : die Standardabweichungen von  $X$  und  $Y$

$\kappa_{XY}$ : der Korrelationskoeffizient von  $X$  und  $Y$

Wir werden sehen:

$\kappa^2$  ist ein Maß dafür, um wieviel besser man  $Y$  durch eine affin lineare Funktion von  $X$  vorhersagen kann:

$$Y = \beta_1 X + \beta_0 + \text{“Fehler”},$$

als durch eine Konstante:

$$Y = c + \text{“Fehler”}.$$

(Die “Güte der Vorhersage” bezieht sich auf die Kleinheit des **erwarteten quadratischen Fehler (mean square error)**.)

## 5. Beste konstante Vorhersage

Um die eben behauptete Eigenschaft von  $\kappa^2$  einzusehen,  
fragen wir erst einmal:

Durch welche **Konstante** wird die Zufallsvariable  $Y$   
(im Sinn des erwarteten quadratischen Fehlers)  
am besten vorhergesagt?

Durch ihren **Erwartungswert  $E[Y]$**  !

Denn:

$$\begin{aligned}\mathbf{E}[(Y - c)^2] &= \mathbf{E}[(Y - \mu_Y + \mu_Y - c)^2] \\ &= \mathbf{E}[(Y - \mu_Y)^2] + 2\mathbf{E}[(Y - \mu_Y)(\mu_Y - c)] + (\mu_Y - c)^2 \\ &= \sigma_Y^2 + 0 + (\mu_Y - c)^2.\end{aligned}$$

Das wird minimiert von

$$c = \mu_Y$$

und hat den Minimalwert

$$\sigma_Y^2.$$

## 6. Beste affin lineare Vorhersage

Durch welche **affin lineare Funktion** von  $X$ ,

$$\beta_1 X + \beta_0,$$

wird die Zufallsvariable  $Y$

(wieder im Sinn des erwarteten quadratischen Fehlers)

am besten vorhergesagt?

Genauer:

Für welche Zahlen  $\beta_1, \beta_0$  wird  
 $\mathbf{E}[(Y - \beta_1 X - \beta_0)^2]$  minimal?

Wie wir gleich sehen werden, ist die Lösung:

$$\beta_1 := \frac{\sigma_Y}{\sigma_X} \kappa_{XY}$$

und  $\beta_0$  so, dass  $\mu_Y = \beta_1 \mu_X + \beta_0$ .

M. a. W.:  $\beta_0$  so, dass der Punkt  $(\mu_X, \mu_Y)$   
auf der Geraden  $y = \beta_1 x + \beta_0$  liegt.

Wir nennen diese Gerade  
die **Regressionsgerade** für  $Y$  auf der Basis von  $X$ .

Wir begründen jetzt die Behauptung über  $\beta_0$  und  $\beta_1$ :

$$\begin{aligned} & \mathbf{E}[(Y - \beta_1 X - \beta_0)^2] \\ &= \mathbf{Var}[Y - \beta_1 X - \beta_0] + (\mathbf{E}[Y - \beta_1 X - \beta_0])^2 \\ &= \mathbf{Var}[Y - \beta_1 X] + (\mu_Y - \beta_1 \mu_X - \beta_0)^2 \end{aligned}$$

Der **zweite Summand** ist Null für  $\beta_0 = \mu_Y - \beta_1 \mu_X$ .

Damit haben wir schon mal **die eine Bedingung** gefunden.

Für welches  $\beta_1$  wird **der erste Summand** minimal?

Im Rest des Abschnittes schreiben wir um der besseren Lesbarkeit willen  $\kappa$  statt  $\kappa_{XY}$ .

$$\mathbf{Var}[Y - \beta_1 X] = \mathbf{Var}Y - 2\beta_1 \mathbf{Cov}[X, Y] + \beta_1^2 \mathbf{Var}X$$

$$= \sigma_Y^2 - 2\beta_1 \kappa \sigma_X \sigma_Y + \beta_1^2 \sigma_X^2$$

$$= \sigma_Y^2 - \sigma_Y^2 \kappa^2 + (\sigma_Y^2 \kappa^2 - 2\beta_1 \kappa \sigma_X \sigma_Y + \beta_1^2 \sigma_X^2)$$

$$\mathbf{Var}[Y - \beta_1 X] = \mathbf{Var}Y - 2\beta_1 \mathbf{Cov}[X, Y] + \beta_1^2 \mathbf{Var}X$$

$$= \sigma_Y^2 - 2\beta_1 \kappa \sigma_X \sigma_Y + \beta_1^2 \sigma_X^2$$

$$= \sigma_Y^2 - \sigma_Y^2 \kappa^2 + (\sigma_Y \kappa - \beta_1 \sigma_X)^2$$

Der rechte Summand wird Null für

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \kappa.$$

Und der Minimalwert von  $\mathbf{Var}[Y - \beta_1 X]$  ist  $\sigma_Y^2(1 - \kappa^2)$ .

Damit ist auch der Minimalwert von  $\text{Var}[Y - \beta_1 X - \beta_0 \mathbf{1}]$  gleich  $\sigma_Y^2(1 - \kappa^2)$ .

Der Minimalwert von  $\text{Var}[Y - c \mathbf{1}]$  war  $\sigma_Y^2$ .

Die Verbesserung der Approximation (“Vorhersage”) von  $Y$  im quadratischen Mittel, wenn man zu den Vielfachen von 1

die Vielfachen von  $X$  dazunimmt, beträgt

$$\sigma_Y^2 - \sigma_Y^2(1 - \kappa^2) = \kappa^2 \sigma_Y^2.$$

Also ist der Anteil von  $\sigma_Y^2$ ,

der von den Vielfachen von  $X$  zusätzlich zu den Vielfachen von 1 “erklärt” wird, gleich  $\kappa^2 \sigma_Y^2$ .

Wir halten fest: Die Minimierungsaufgabe

$$\mathbf{E}[(Y - \beta_1 X - \beta_0)^2] \stackrel{!}{=} \min$$

für die *beste affin lineare Vorhersage von Y*

*auf der Basis von X*

(im Sinn des quadratischen Mittels)

hat die Lösung

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \kappa, \quad \mu_Y = \beta_1 \mu_X + \beta_0$$

und den Minimalwert  $(1 - \kappa^2) \sigma_Y^2$ .

## 7. Beispiel:

Gemeinsam normalverteilte Zufallsvariable

$Z_1, Z_2$  seien unabhängig und standard-normalverteilt,

Wir wählen eine Konstante  $\kappa \in [-1, 1]$  und setzen

$$X := Z_1, \quad Y := \kappa Z_1 + \sqrt{1 - \kappa^2} Z_2.$$

Damit ergibt sich  $\sigma_X^2 = \sigma_Y^2 = 1$ ,

$$\text{Cov}[X, Y] = \text{Cov}[Z_1, \rho Z_1] = \kappa.$$

Somit gilt hier:

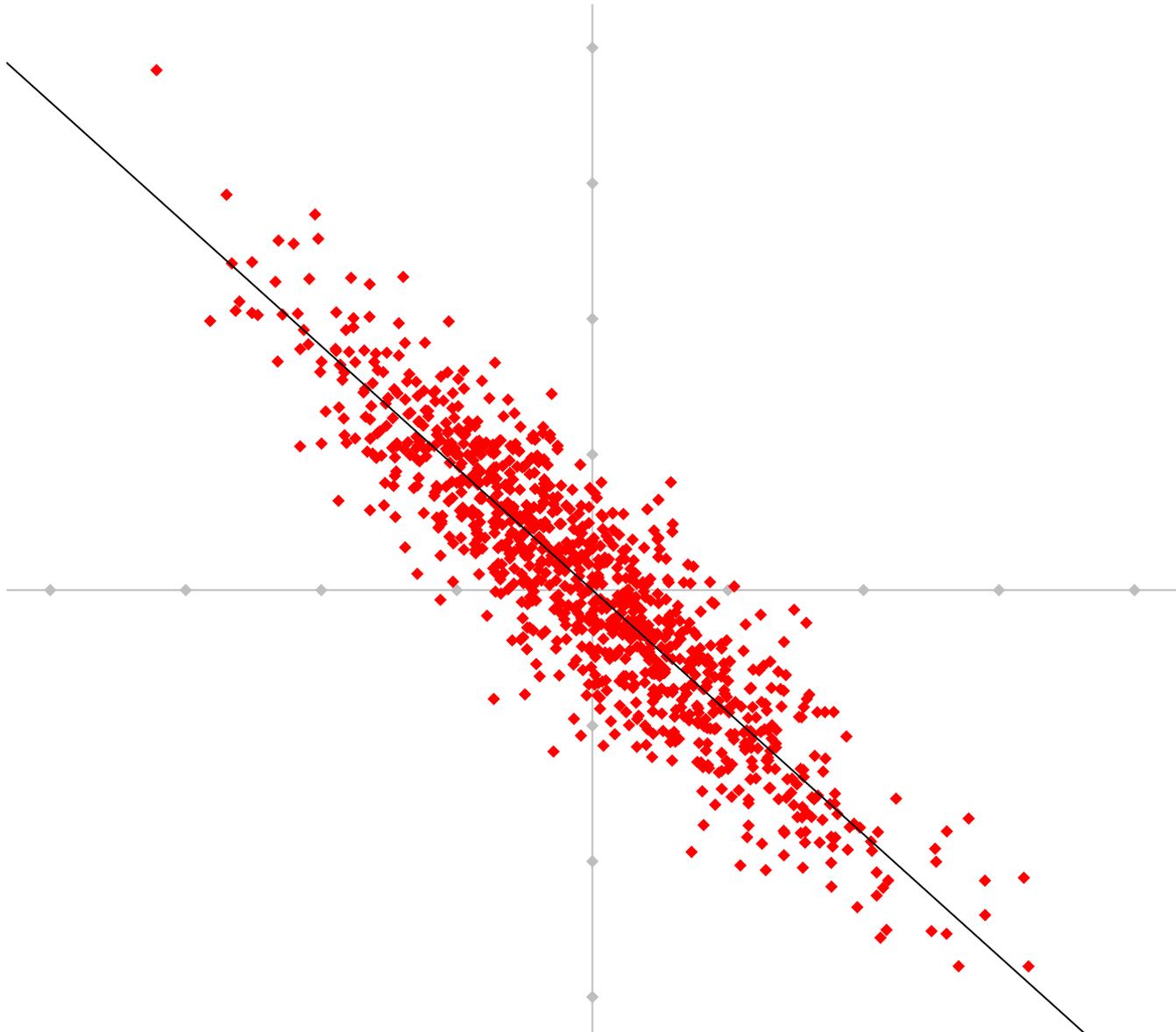
$$\kappa_{XY} = \kappa$$

(die Bezeichnung der Konstanten war hier also mit Bedacht gewählt).

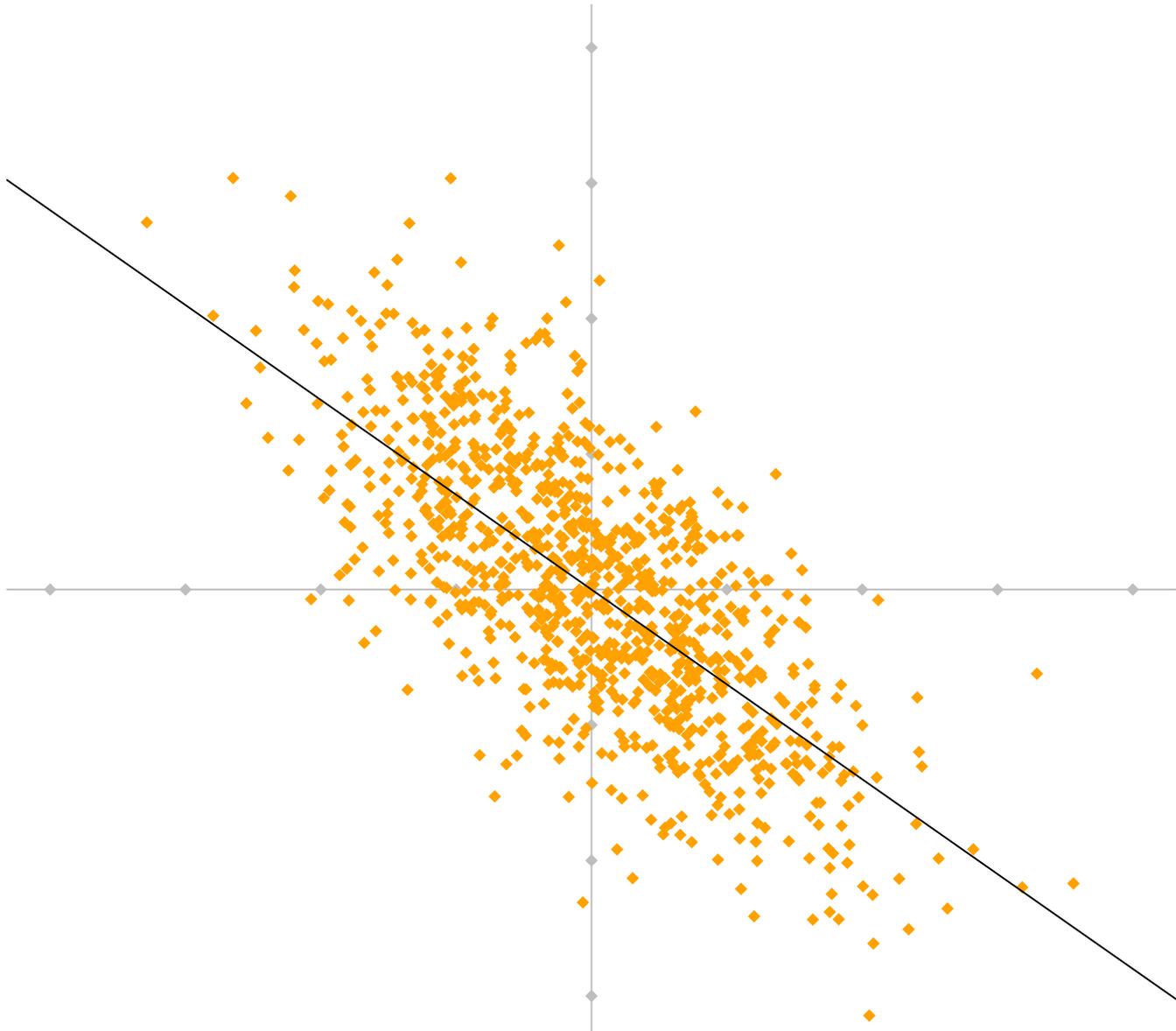
Auch  $Y$  ist übrigens standard-normalverteilt (s. V6b Folie 58).

Die folgenden Bilder  
( $\kappa = -0.9, -0.7, \dots, 0.7, 0.9$ )  
zeigen jeweils die Realisierungen von  
1000 unabhängige Kopien  $(X_i, Y_i)$  von  $(X, Y)$ ,  
zusammen mit der  
Regressionsgeraden für  $Y$  auf der Basis von  $X$

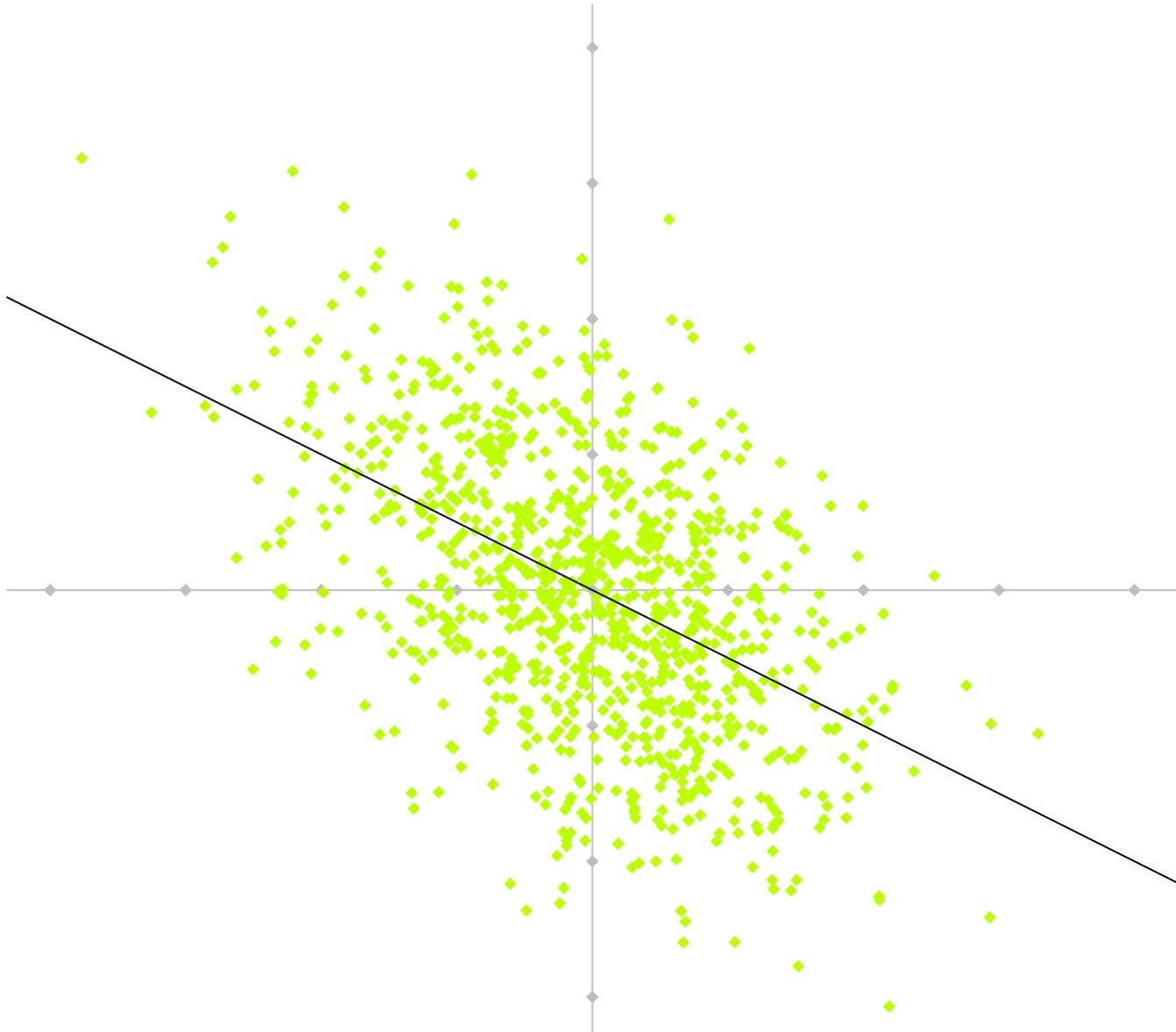
Korrelation = - 0.9



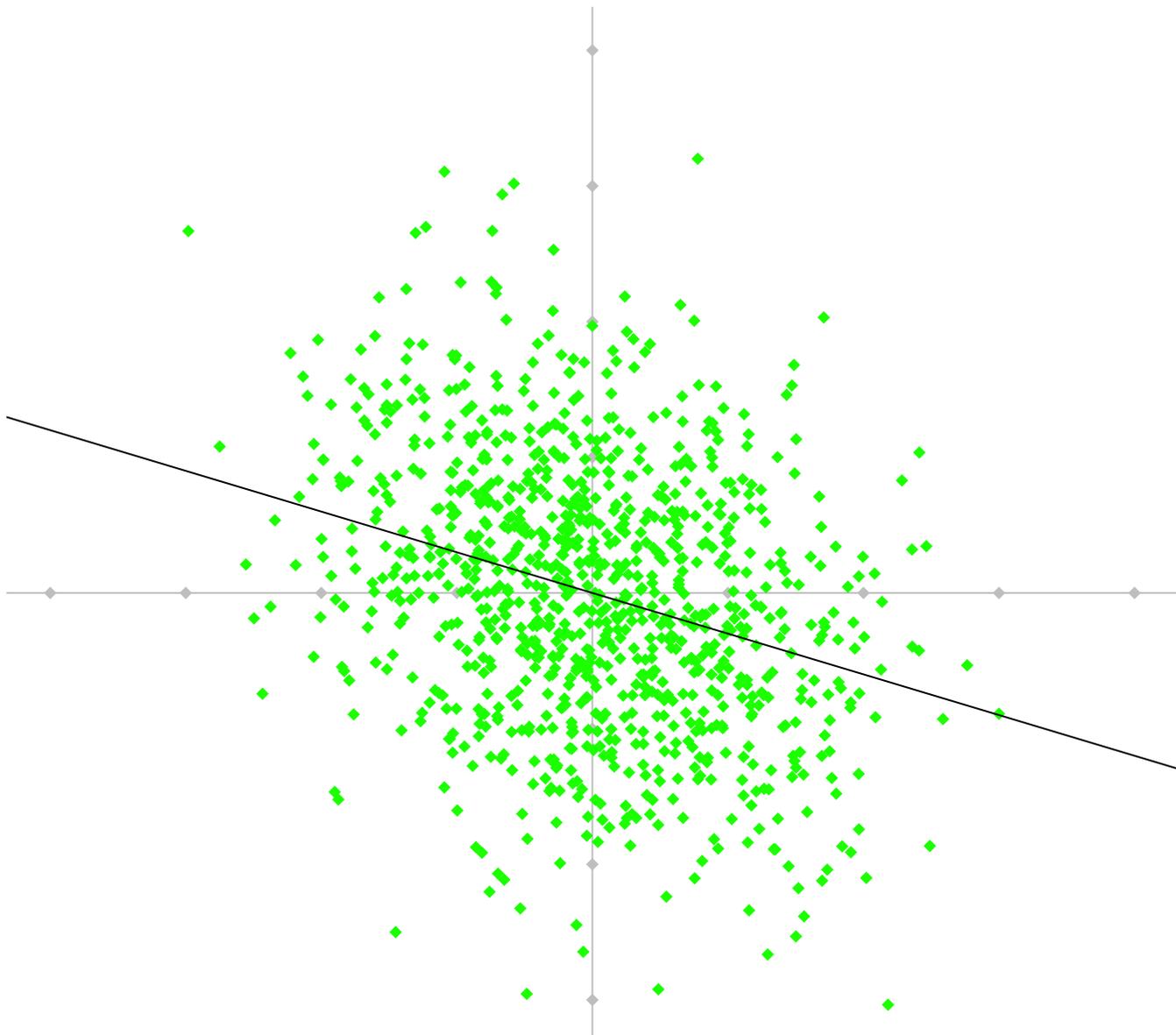
Korrelation = - 0.7



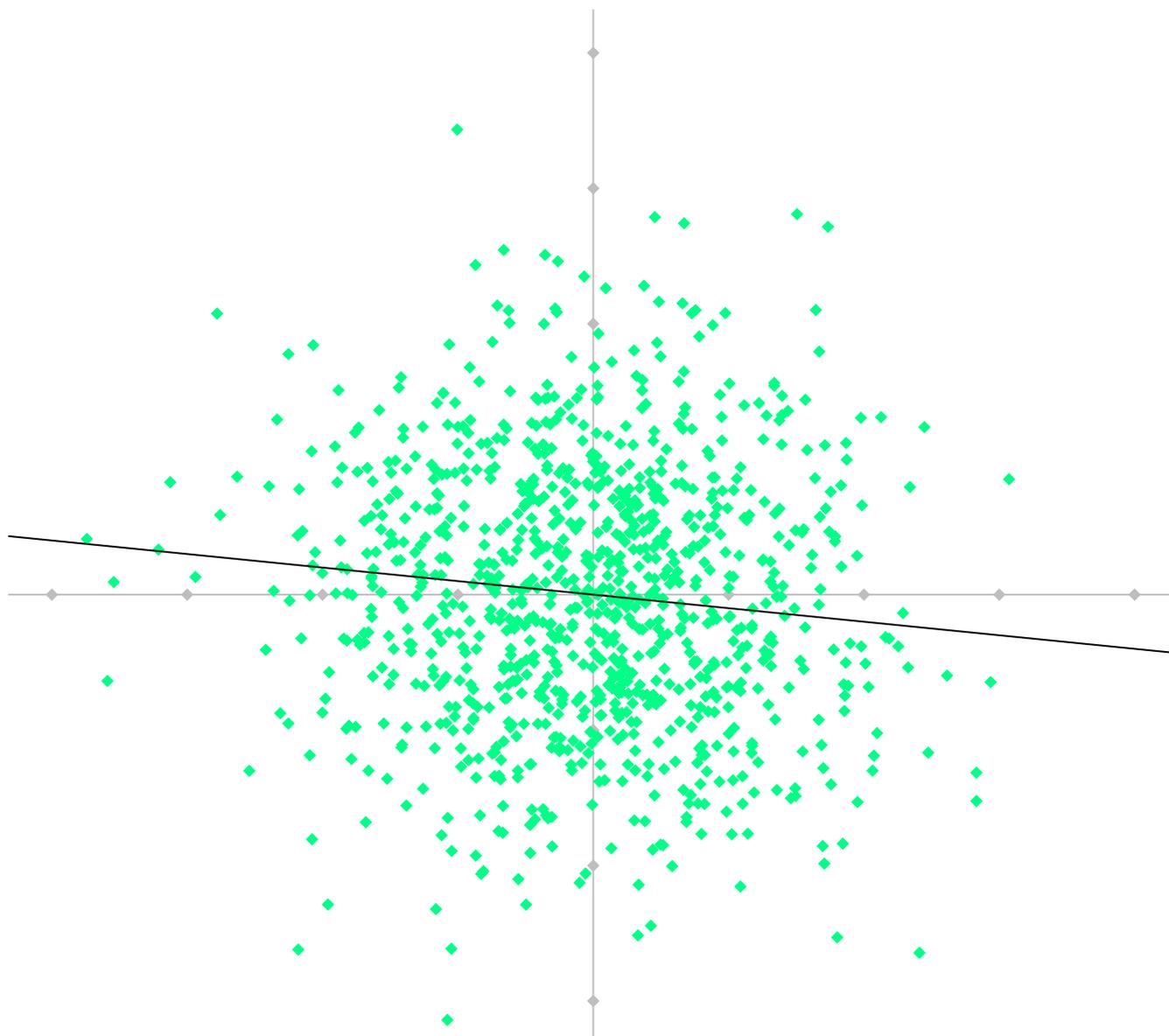
Korrelation = - 0.5



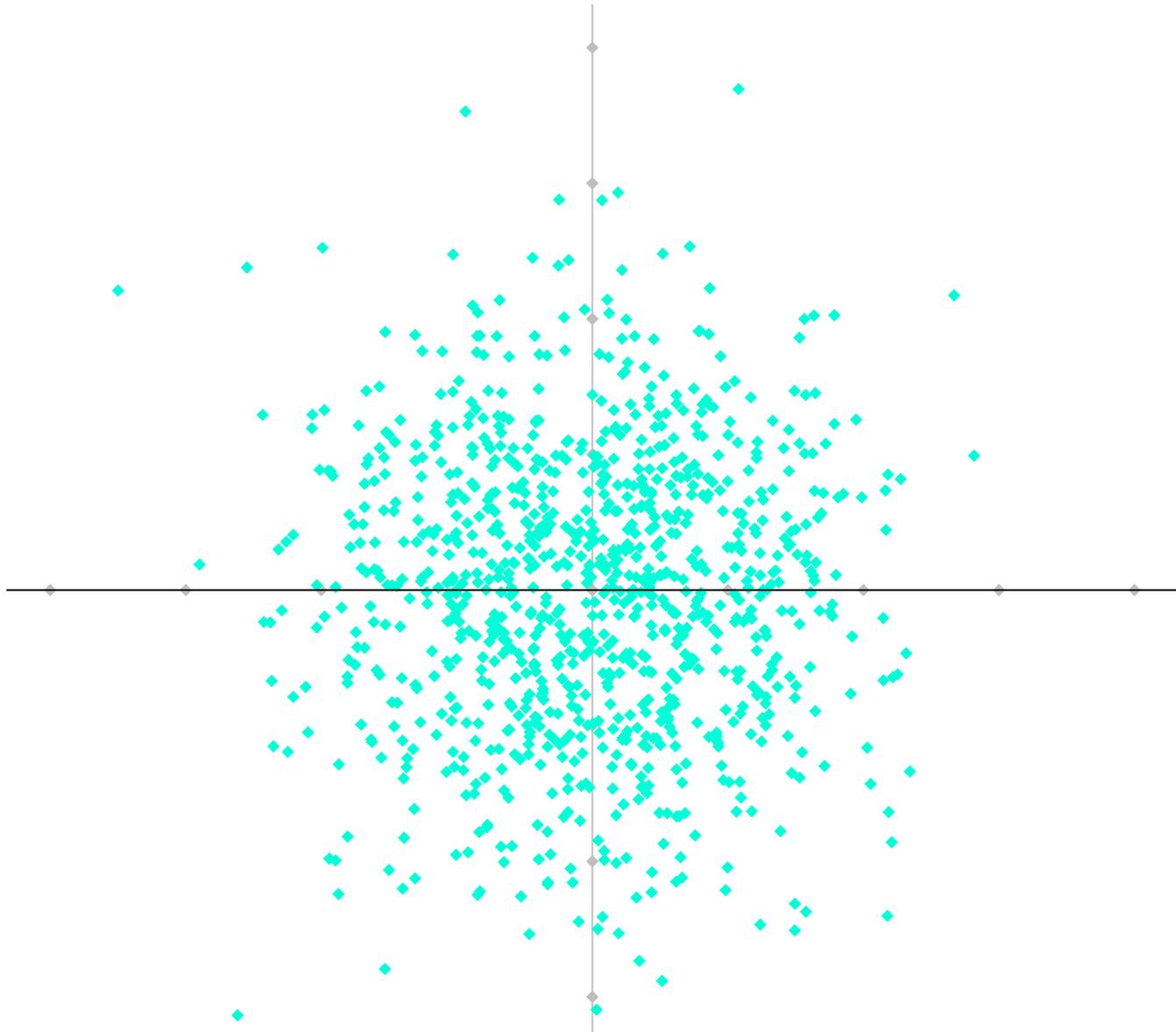
Korrelation = - 0.3



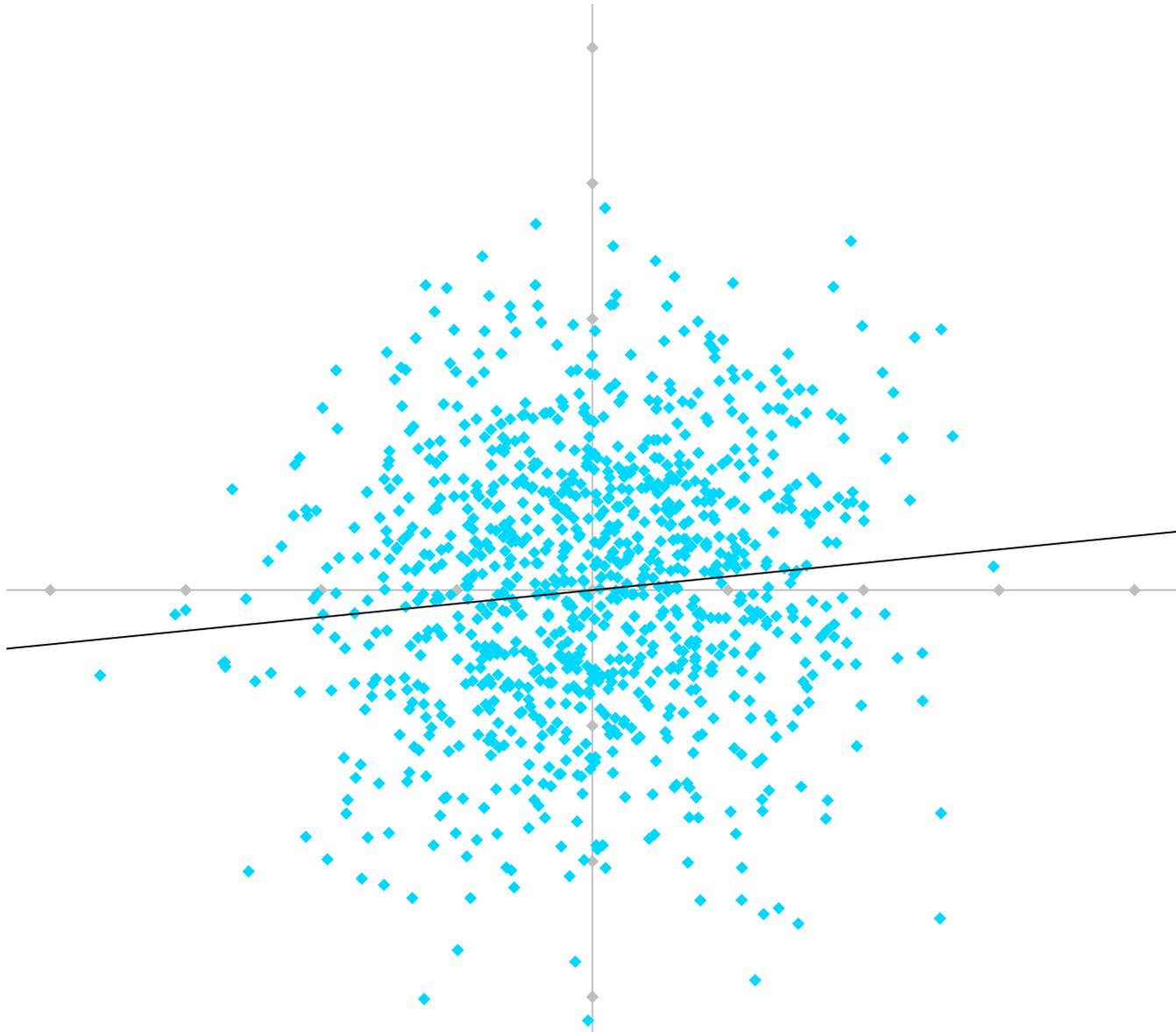
Korrelation = - 0.1



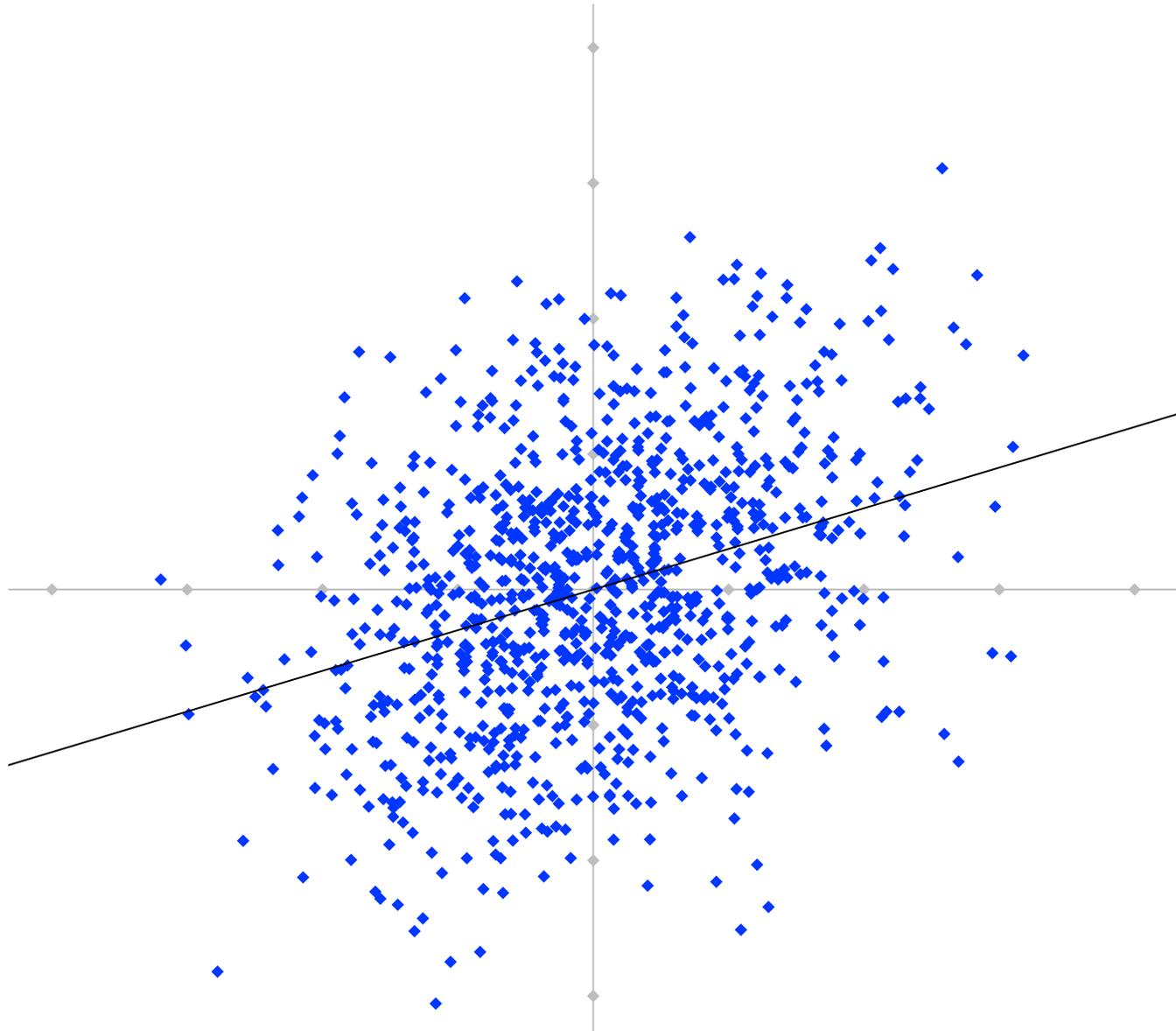
Korrelation = 0



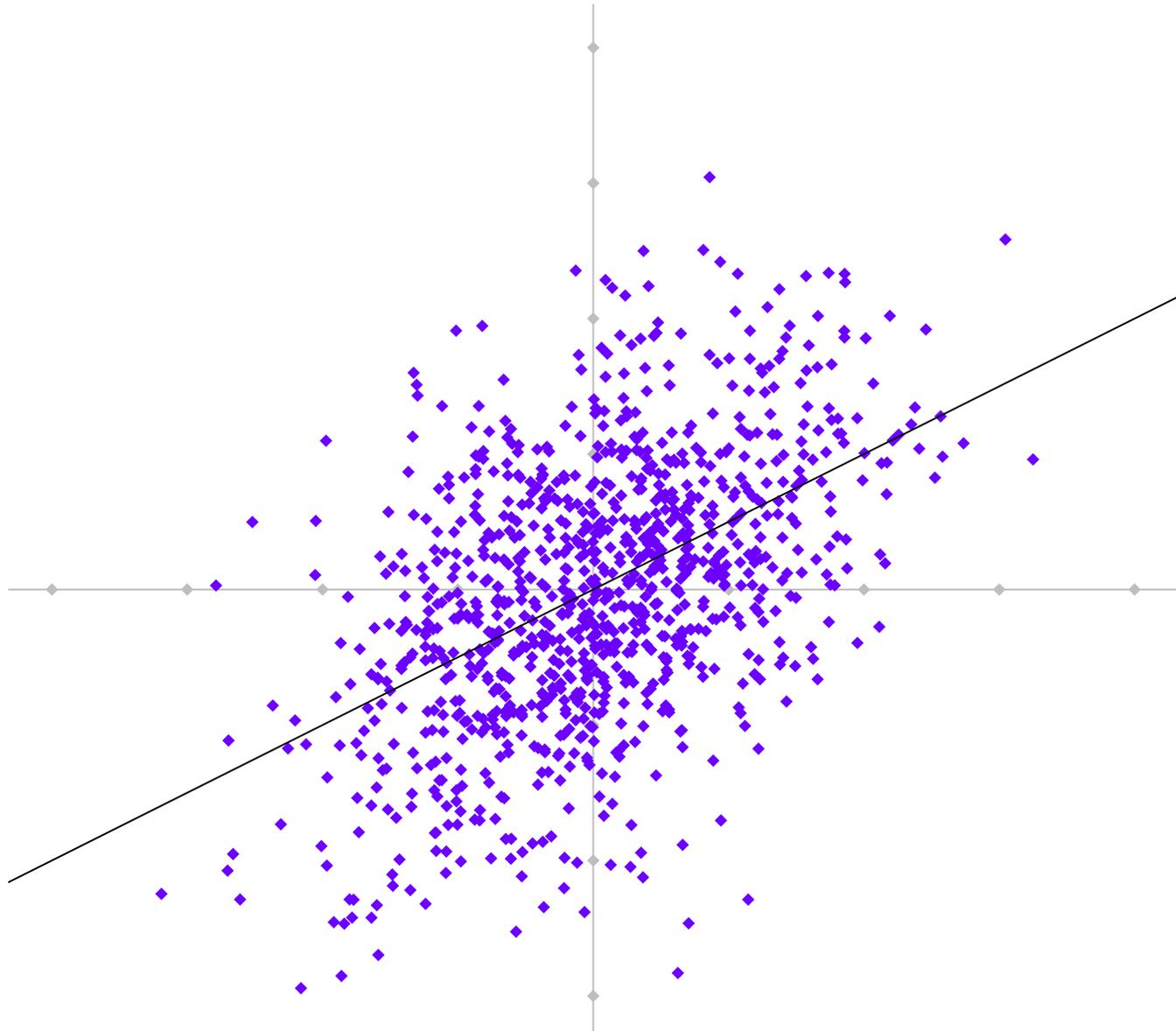
Korrelation = 0.1



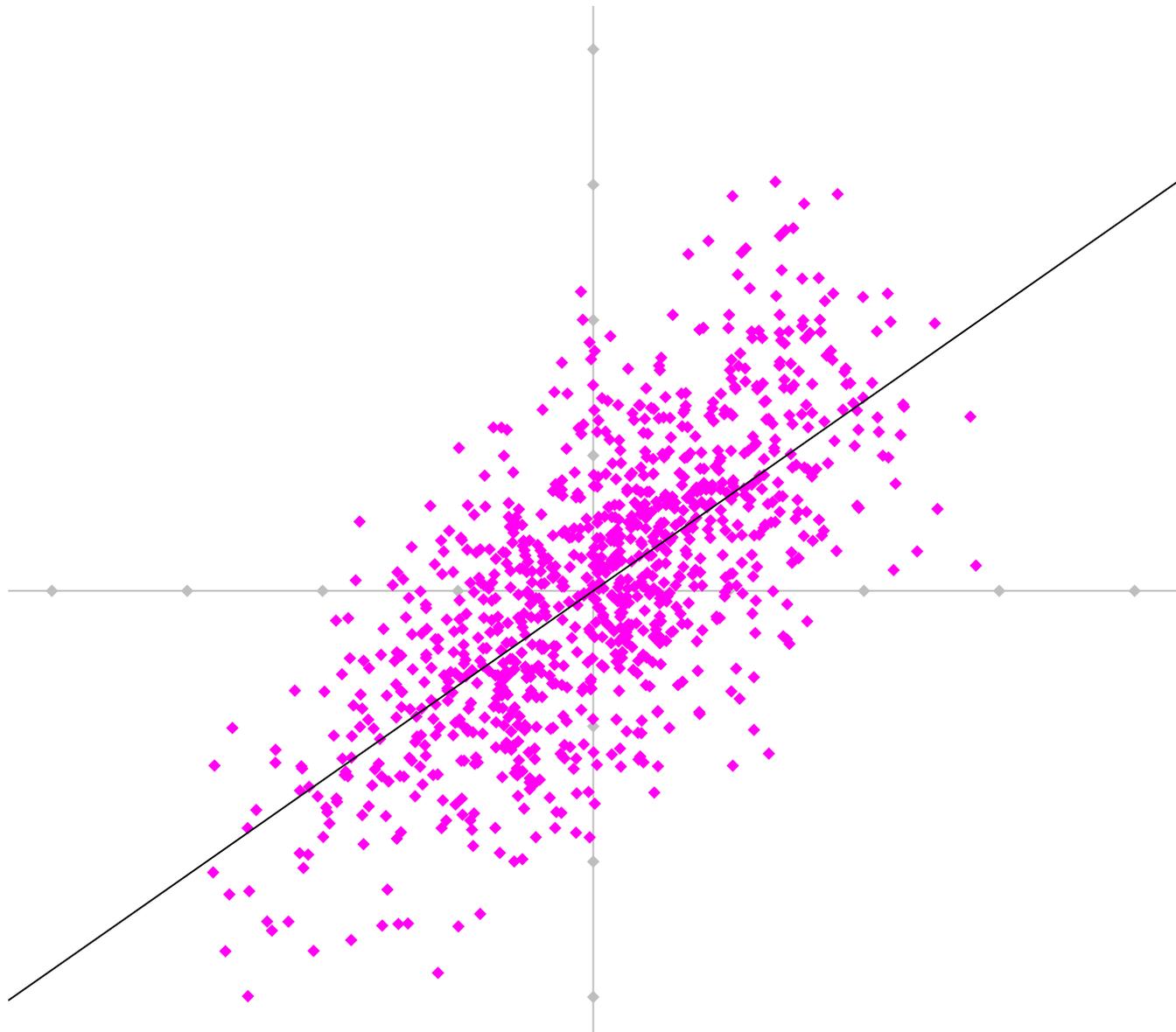
Korrelation = 0.3



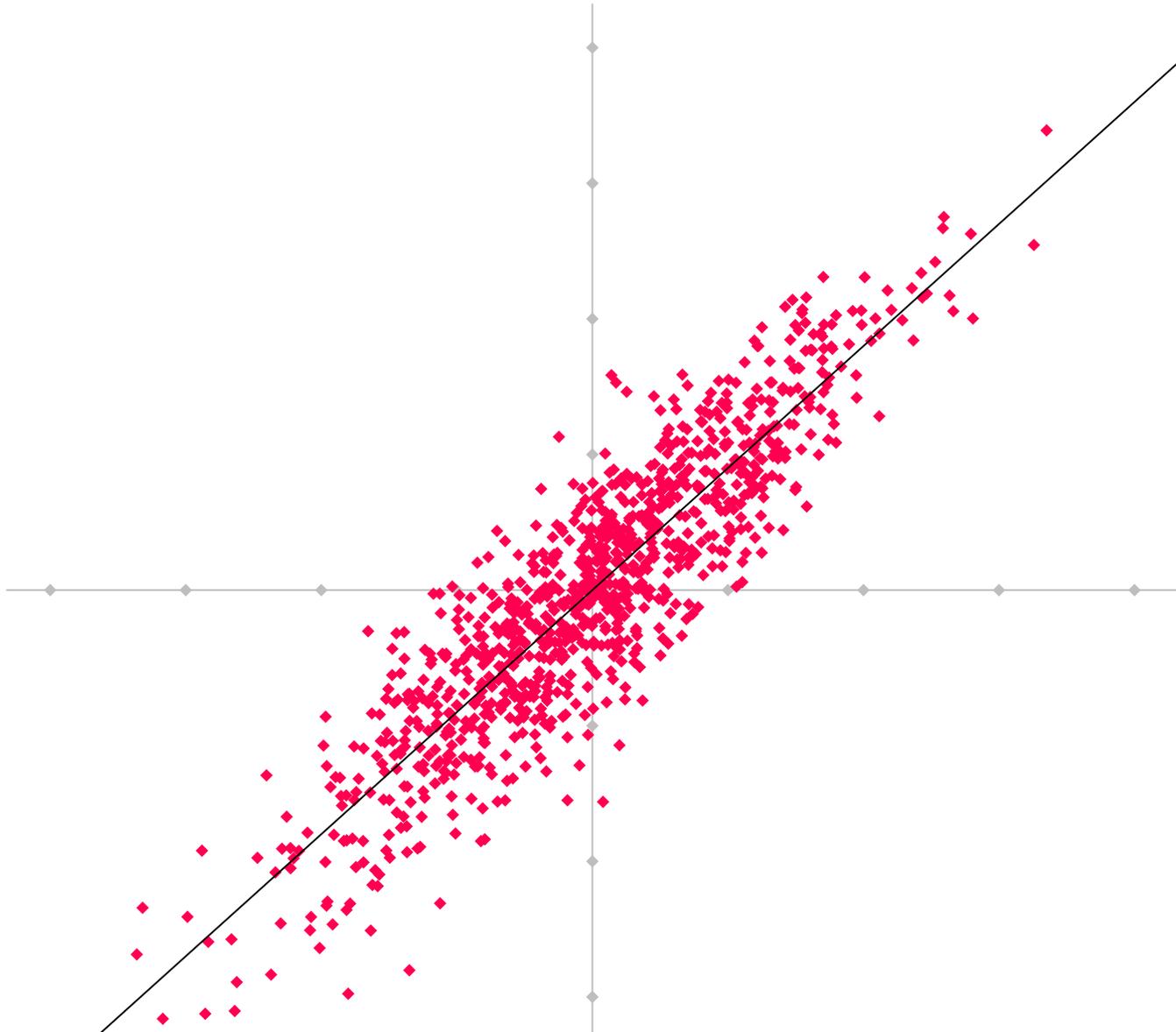
Korrelation = 0.5



Korrelation = 0.7



Korrelation = 0.9



8. Beispiel: “Welche Gerade passt am besten?”

$(x_1, y_1), \dots, (x_n, y_n)$  seien  $n$  verschiedene Punkte im  $\mathbb{R}^2$ .

$(X, Y)$  sei eine rein zufällige Wahl daraus:

$$\mathbf{P}((X, Y) = (x_i, y_i)) = \frac{1}{n}, \quad i = 1, \dots, n.$$

Dann ist

$$\mathbf{E}X = \frac{1}{n} \sum x_i =: \bar{x}$$

$$\sigma_X^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\mathbf{Cov}[X, Y] = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\kappa := \kappa_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}.$$

$$\mathbf{E}[(Y - \beta_1 X - \beta_0)^2] = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)^2$$

wird, wie wir gezeigt haben, minimiert durch

$$\beta_1 := \frac{\sigma_Y}{\sigma_X} \kappa = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

und  $\beta_0$  so, dass  $\bar{y} = \beta_1 \bar{x} + \beta_0$ .

Diese Gerade  $y = \beta_1 x + \beta_0$  heißt die **Regressionsgerade** zu den Punkten  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .  
(oder auch die mit der Methode der kleinsten Quadrate gefundene Ausgleichsgerade).

Eine anschauliche Illustration des Themas “Korrelation und Regression” finden Sie auf von Brooks Ferebee konzipierten [Folien](#) zur Vorlesung “Statistik für Biologen”