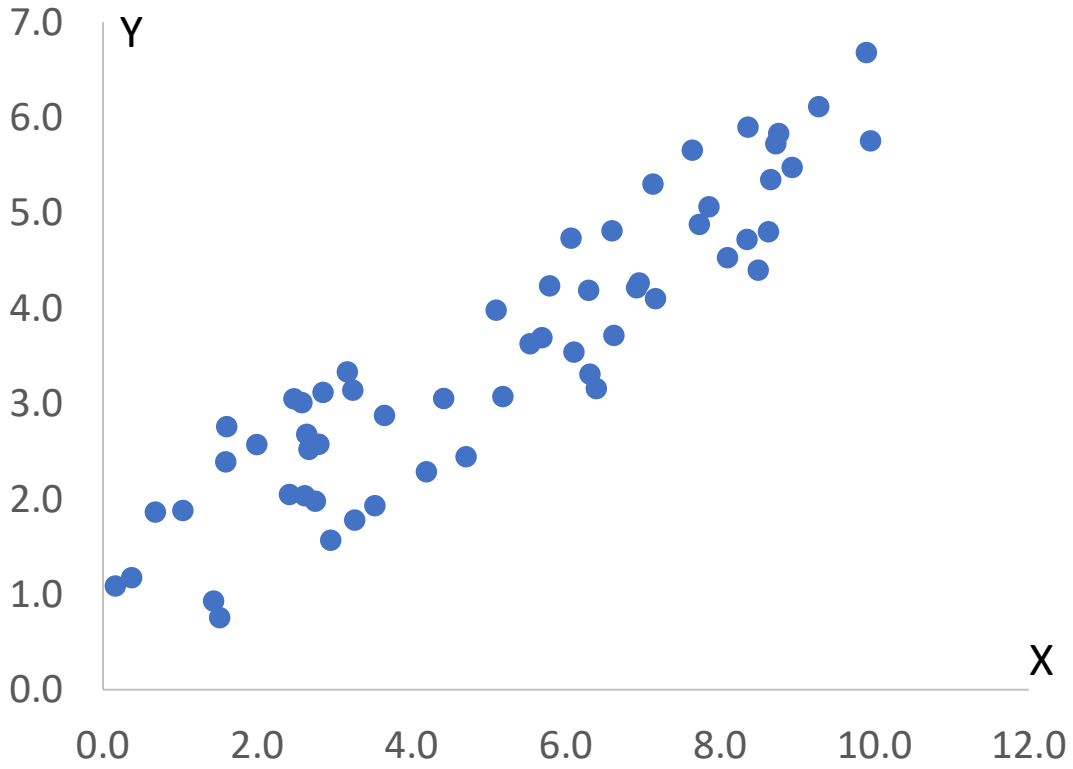
Krótko o hipotezach, testach, ekonometrii

Seminaria Teatr, Muzyka, Cyfryzacja 2022/23

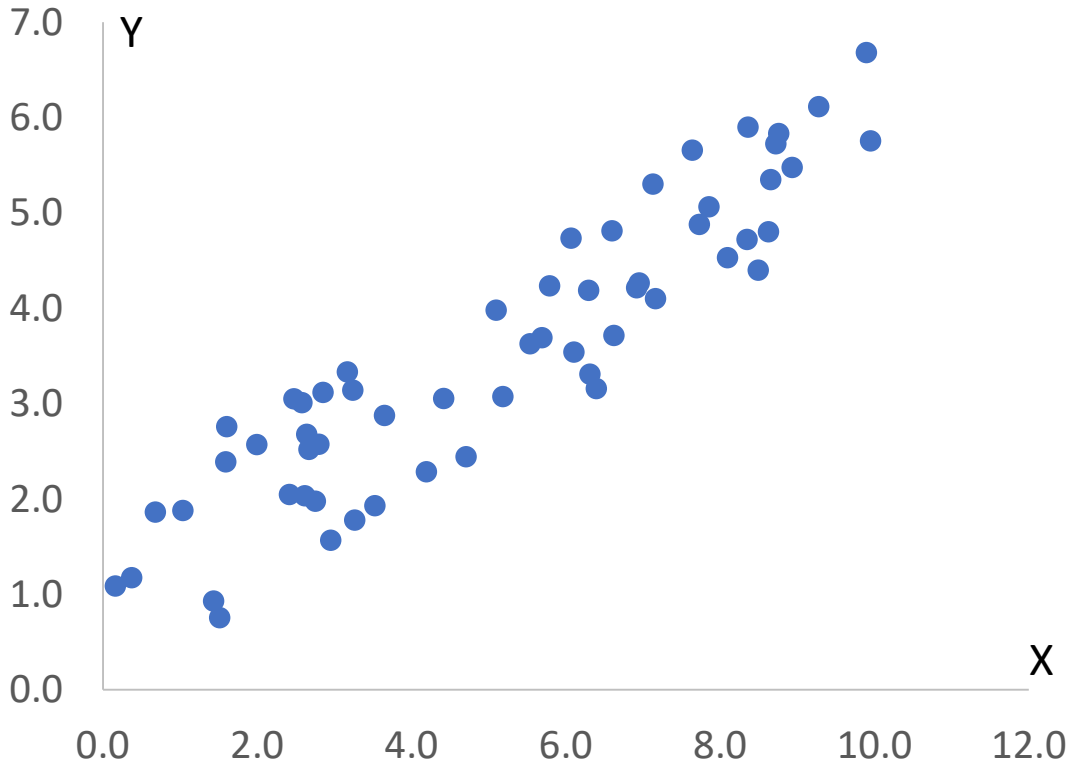
Liniowa zależność – dwie zmienne

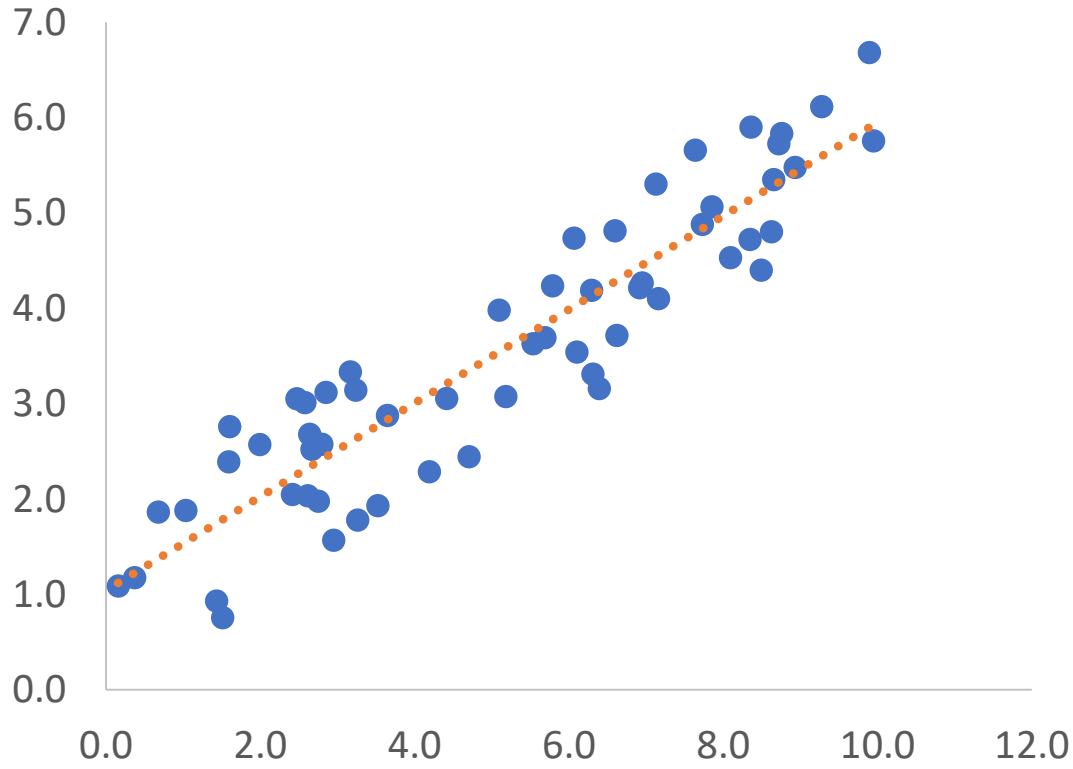
Modelujemy zależność między dwiema zmiennymi:



Liniowa zależność – dwie zmienne

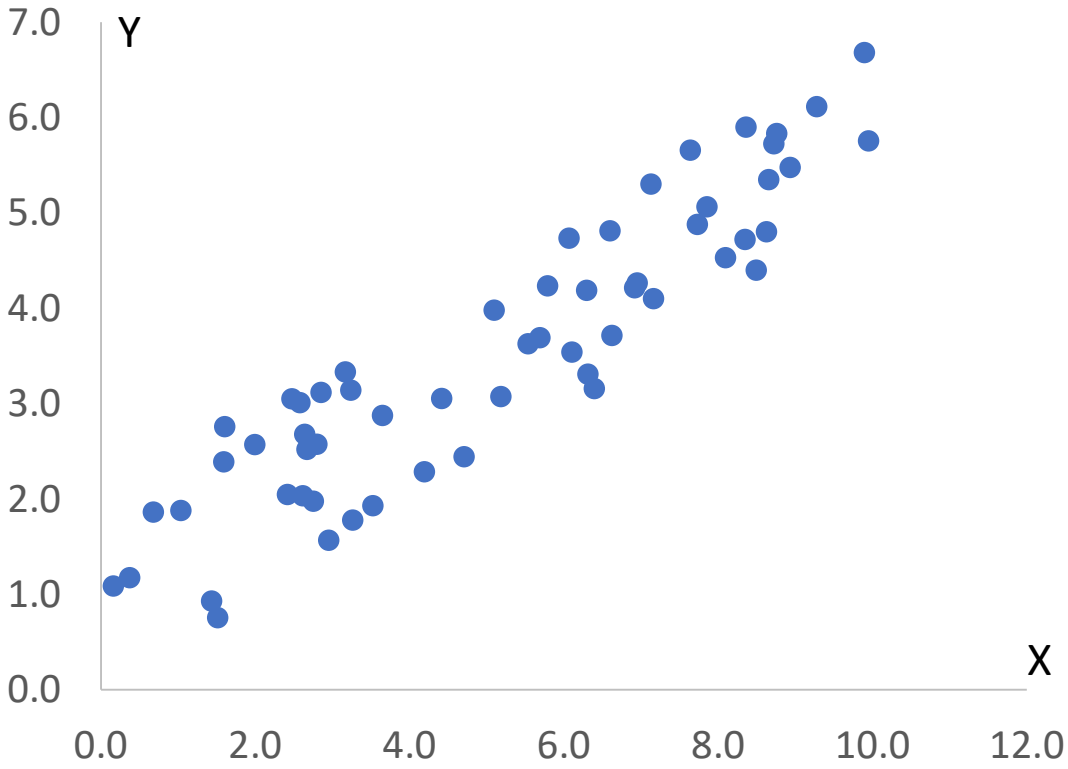
Modelujemy zależność między dwiema zmiennymi: $Y = aX + b$

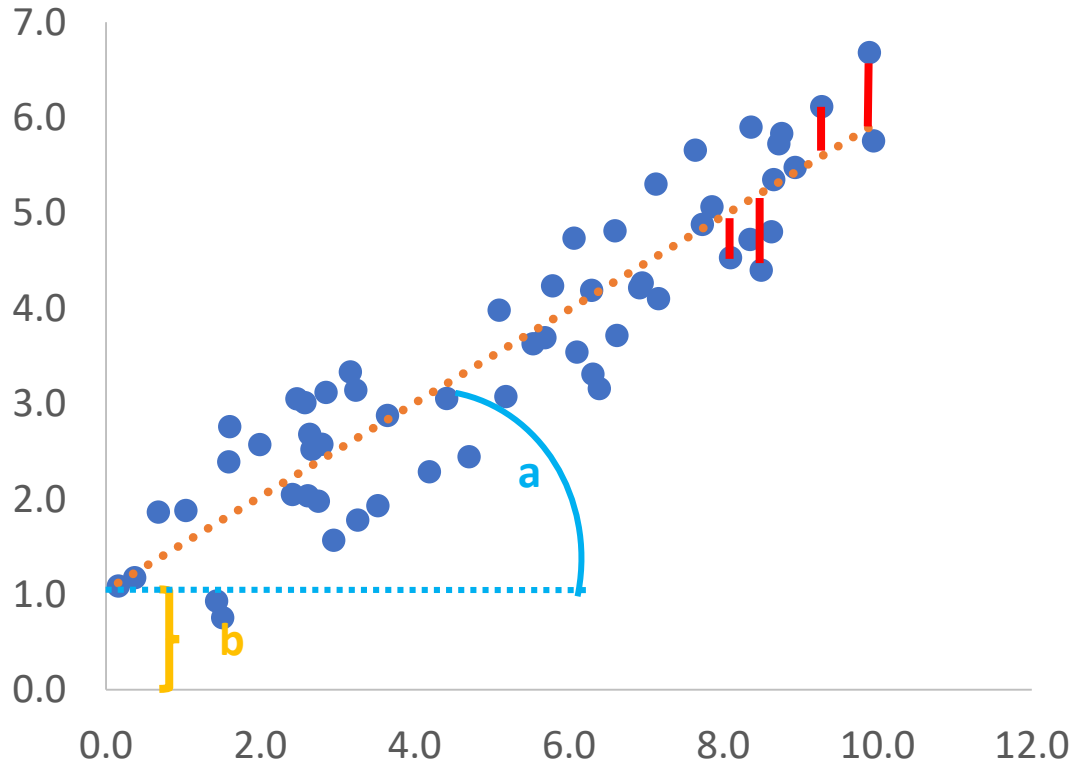




Liniowa zależność – dwie zmienne

Modelujemy zależność między dwiema zmiennymi: Y = aX + b + reszty (błędy) (regresja)





W praktyce zwykle patrzymy na więcej zmiennych

Np. interesuje nas zależność wynagrodzenia i płci:

$$Wynagrodzenie = \beta_0 + \beta_1 Płeć + \epsilon$$

W praktyce zwykle patrzymy na więcej zmiennych

Np. interesuje nas zależność wynagrodzenia i płci:

$$Wynagrodzenie = \beta_0 + \beta_1 Płeć + \epsilon$$

Ale kobiety częściej mają wyższe wykształcenie co wpływa na zarobki.
Możemy chcieć to `wyłączyć':

$$Wynagrodzenie = \beta'_0 + \beta'_1 Płeć + \beta'_2 Wykształcenie + \epsilon'$$

W praktyce zwykle patrzymy na więcej zmiennych

Np. interesuje nas zależność wynagrodzenia i płci:

$$Wynagrodzenie = \beta_0 + \beta_1 Płeć + \epsilon$$

Ale kobiety częściej mają wyższe wykształcenie co wpływa na zarobki.
Możemy chcieć to `wyłączyć':

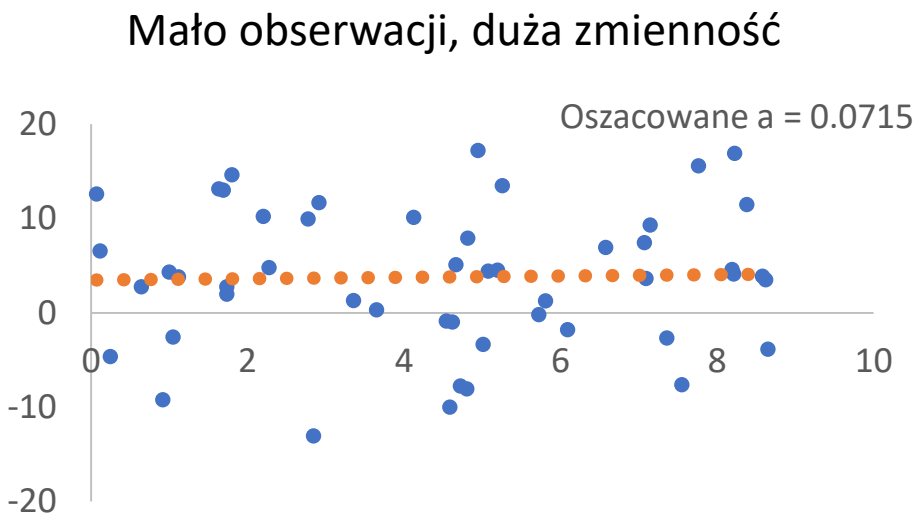
$$Wynagrodzenie = \beta'_0 + \beta'_1 Płeć + \beta'_2 Wykształcenie + \epsilon'$$

A nawet dodać więcej zmiennych, które mogą być ważne by zredukować zmienność:

$$Wynagrodzenie = \beta''_0 + \beta''_1 Płeć + \beta''_2 Wykształcenie + \beta''_3 Wiek + \beta''_4 Staż + \epsilon''$$

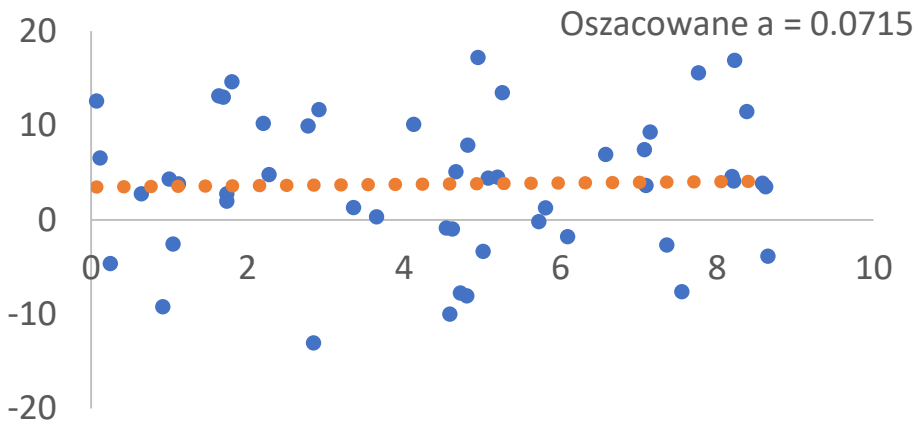
Ale skąd wiadomo czy widziana zależność nie jest dziełem przypadku?

Ta sama zależność między X i Y (a=0.48) i:

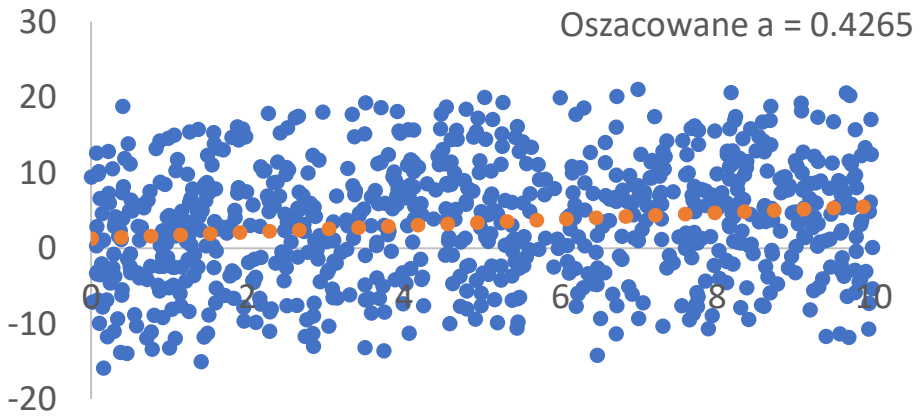


Ta sama zależność między X i Y (a=0.48) i:

Mało obserwacji, duża zmienność

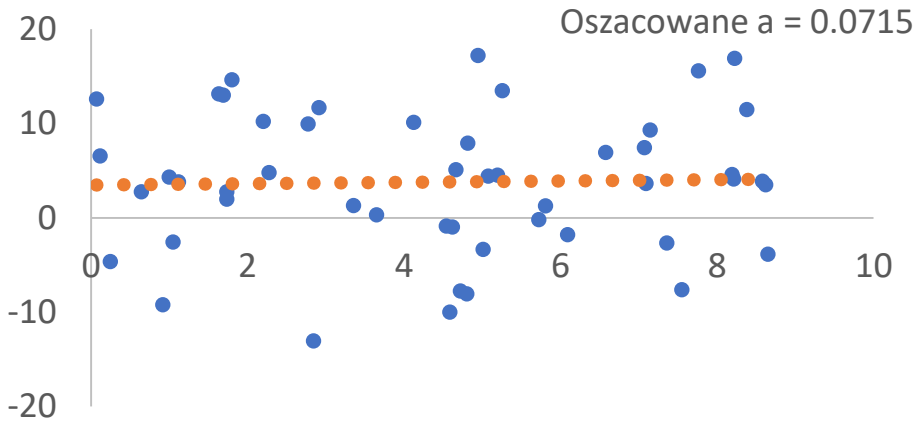


Dużo obserwacji, duża zmienność

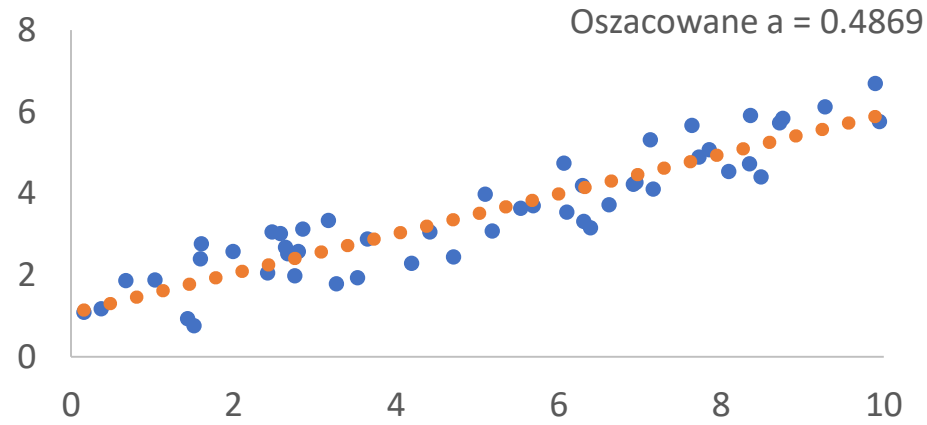


Ta sama zależność między X i Y (a=0.48) i:

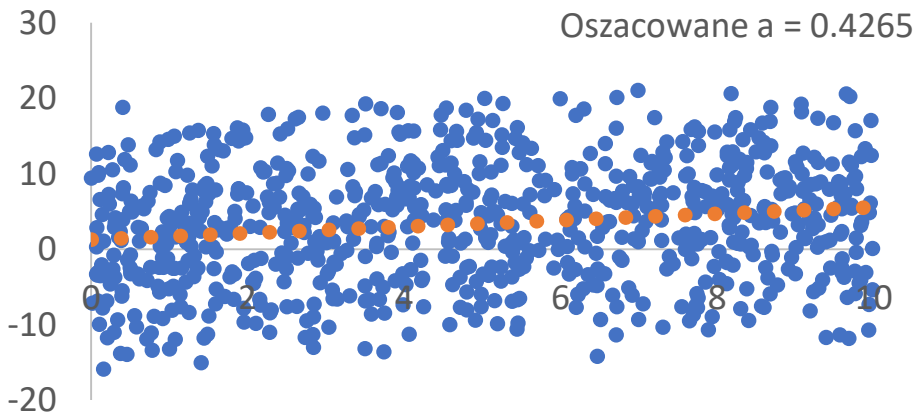
Mało obserwacji, duża zmienność



Mało obserwacji, mała zmienność

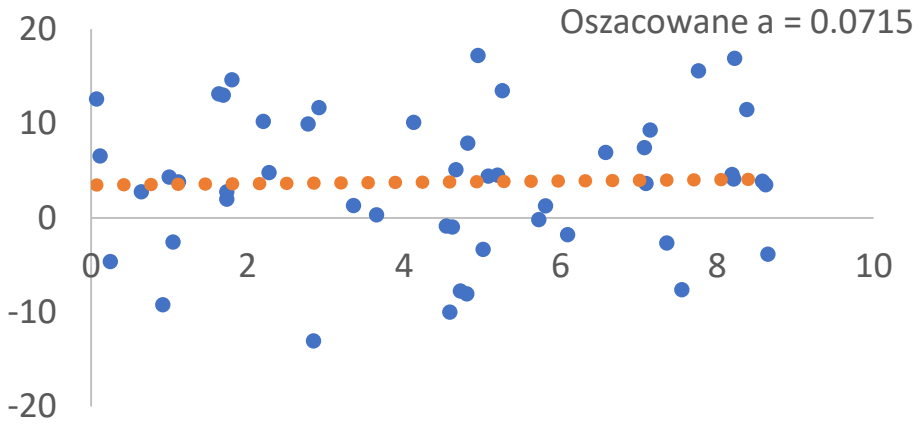


Dużo obserwacji, duża zmienność

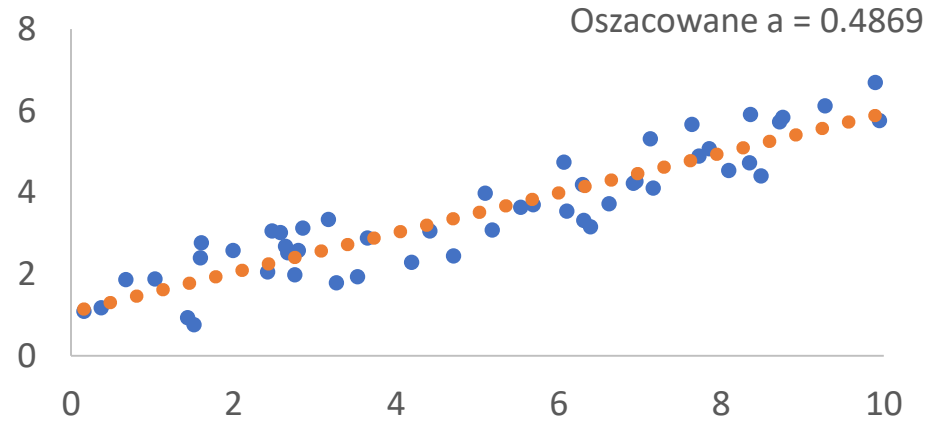


Ta sama zależność między X i Y (a=0.48) i:

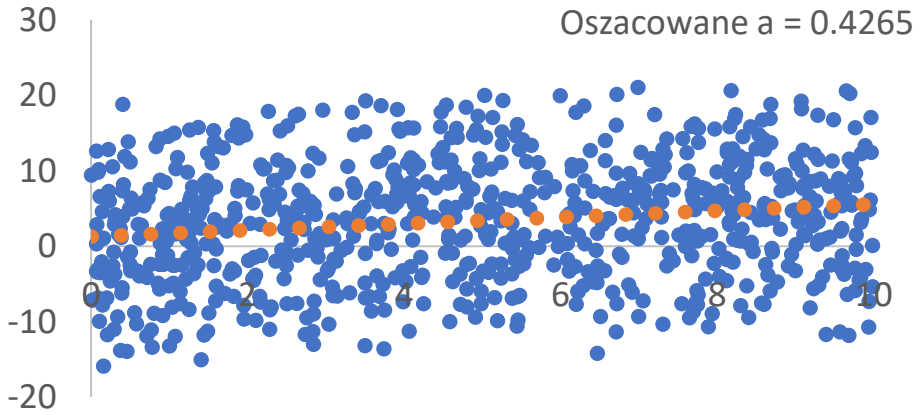
Mało obserwacji, duża zmienność



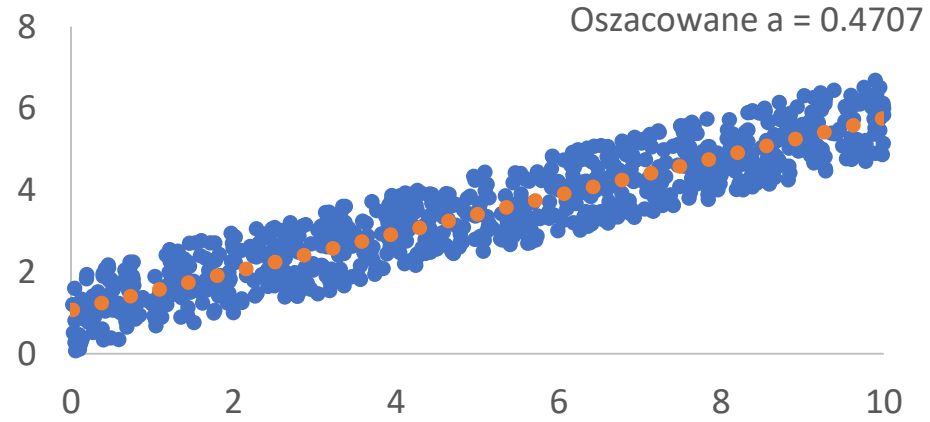
Mało obserwacji, mała zmienność



Dużo obserwacji, duża zmienność



Dużo obserwacji, mała zmienność



Od czego zaczynamy nasze badanie?

Stawiamy hipotezę (lub hipotezy)!

Robiąc badanie często podejrzewamy jakąś zależność między zmiennymi

Stawiamy więc hipotezę:

„Zmienna X jest pozytywnie powiązana ze zmienną Y”

Zwykle stawiamy tę hipotezę jako alternatywę dla tzw. hipotezy zerowej:

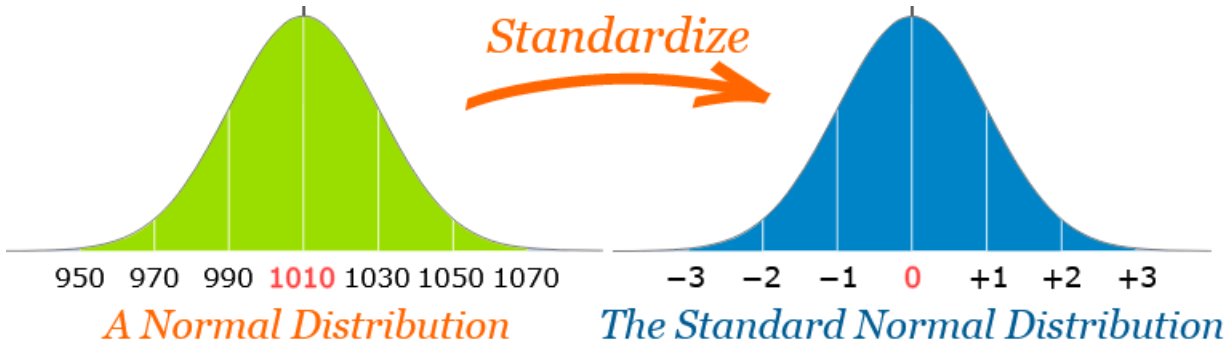
„Nie ma związku między zmienną X i zmienną Y”

Jak decydujemy czy nasza hipoteza została potwierdzona?

- Rachunek prawdopodobieństwa!
- Statystyka!
- Testy!

Idea jest prosta

Centralne Twierdzenie Graniczne pozwala nam sporo sprowadzić do rozkładu Normalnego N(0,1).



Źródło: Mathsuisfun.com: https://www.mathsisfun.com/data/standard-normal-distribution.html (01.11.2022)

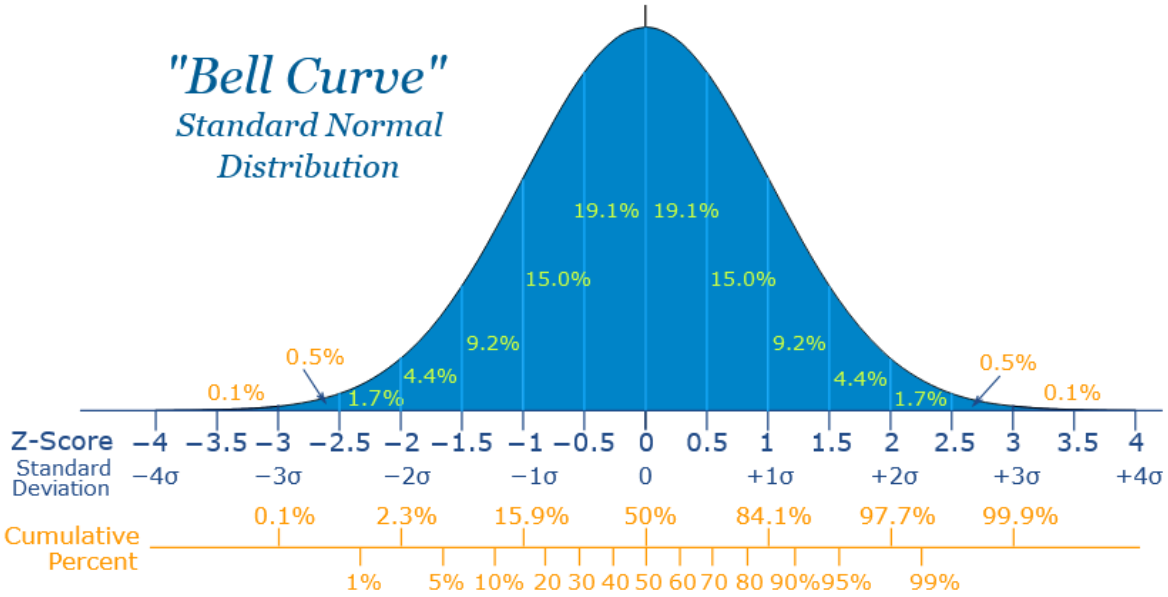
Idea jest prosta

Centralne Twierdzenie Graniczne pozwala nam sporo sprowadzić do rozkładu Normalnego N(0,1).

Dobrze znamy ten rozkład!

Tablice rozkładu normalnego standardowego

x	0	0,01	0,02	0,03	0,04	0,05	0,06
0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147



Źródło: Mathsuisfun.com: https://www.mathsisfun.com/data/standard-normal-distribution.html (01.11.2022)

Idea jest prosta

Centralne Twierdzenie Graniczne pozwala nam sporo sprowadzić do rozkładu Normalnego N(0,1).

Dobrze znamy ten rozkład!

Możemy policzyć statystykę testową tak, że jeśli hipoteza 0 jest prawdziwa, to powinna być z rozkładu N(0,1).

Idea jest prosta

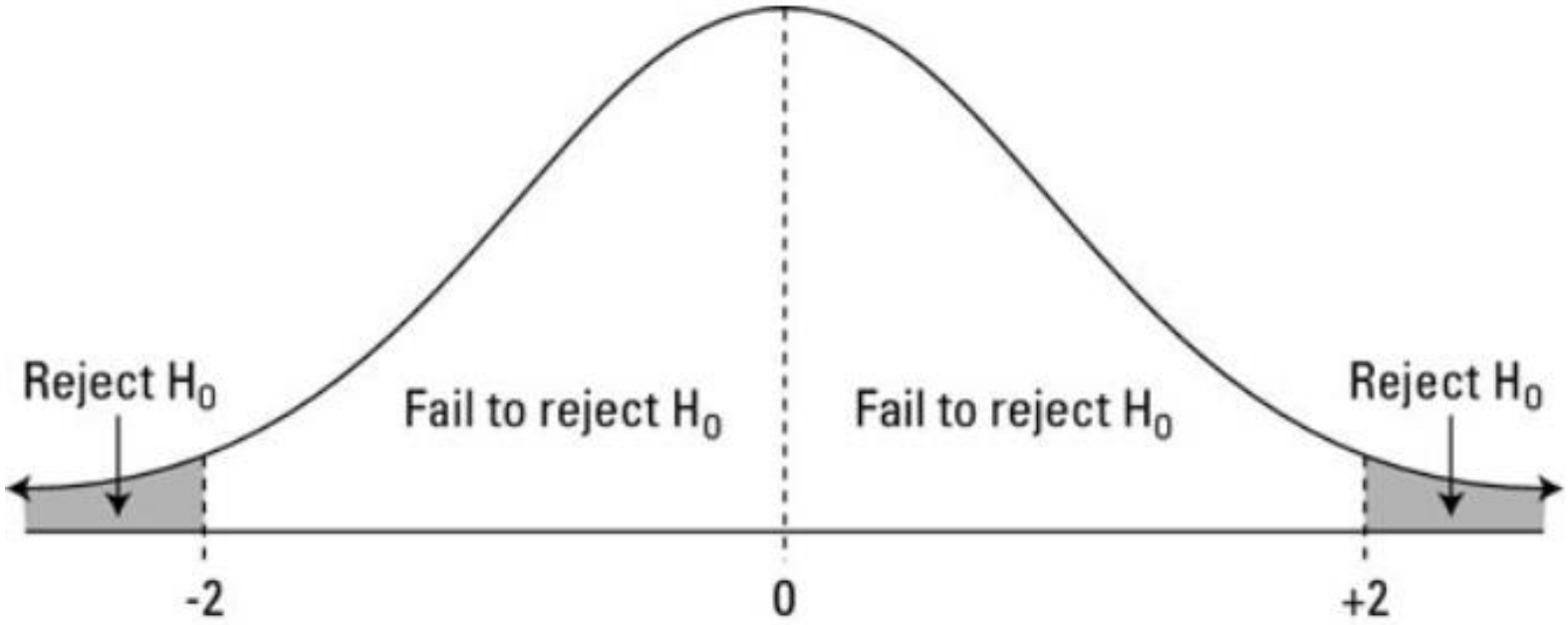
Centralne Twierdzenie Graniczne pozwala nam sporo sprowadzić do rozkładu Normalnego N(0,1).

Dobrze znamy ten rozkład!

Możemy policzyć statystykę testową tak, że jeśli hipoteza 0 jest prawdziwa, to powinna być z rozkładu N(0,1).

Ale jeśli widzimy że marne szanse by była, to odrzucamy H0!

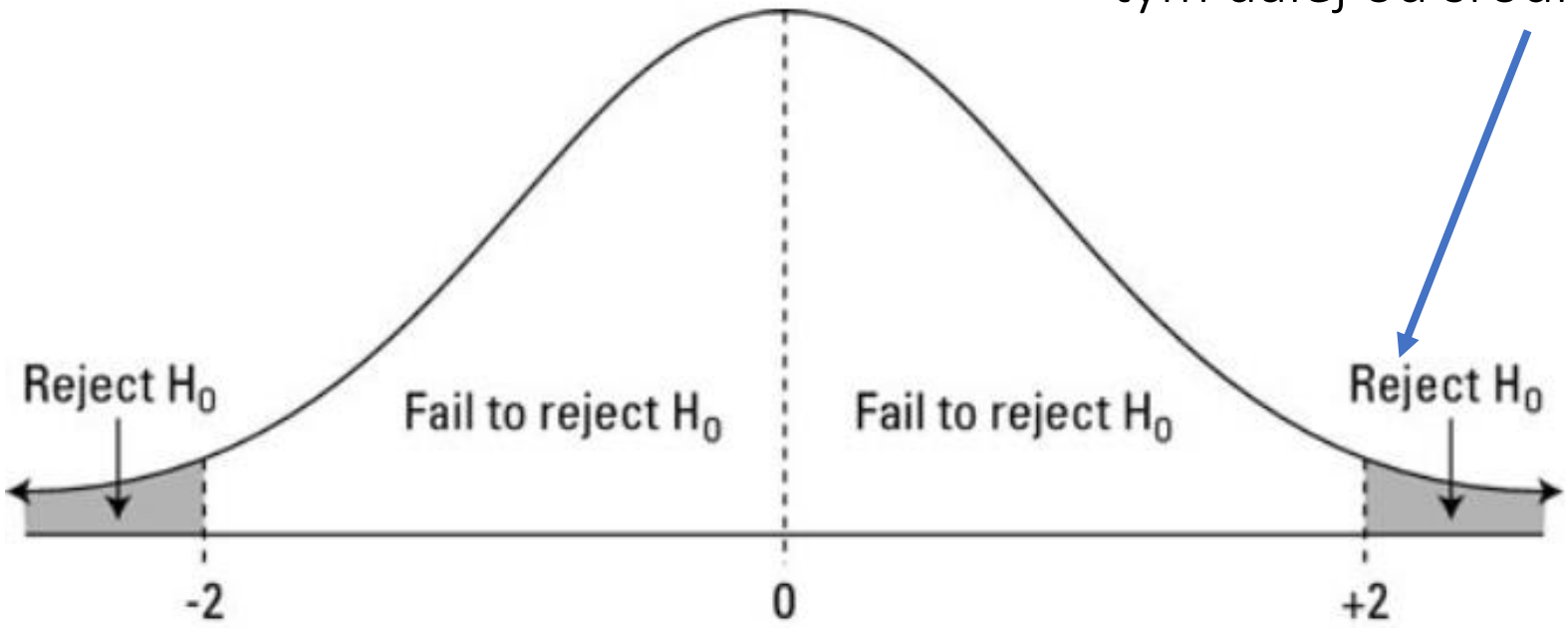
Czyli graficznie:



Źródło: Dummies.com (https://www.dummies.com/article/academics-the-arts/math/statistics/how-to-determine-a-p-value-when-testing-a-null-hypothesis-169062/; 01.11.2022)

Czyli graficznie:

(Ten punkt jest trochę umowny;
im bardziej chcemy być pewni,
tym dalej od środka się znajduje)



Źródło: Dummies.com (https://www.dummies.com/article/academics-the-arts/math/statistics/how-to-determine-a-p-value-when-testing-a-null-hypothesis-169062/; 01.11.2022)

Stata/R/Python większość za nas policzą
Koniec końców mogą nas interesować – na przykład:

- Istotność oszacowania
  (czy udało się potwierdzić hipotezę i jaki jest kierunek?)

- Wielkość oszacowania
  (czy zależność jest duża czy mała?)

Stata/R/Python większość za nas policzą
Koniec końców mogą nas interesować – na przykład:

- Istotność oszacowania
  (czy udało się potwierdzić hipotezę i jaki jest kierunek?)

- Wielkość oszacowania
  (czy zależność jest duża czy mała?)

Zależność może być istotna statystycznie, ale czasem nie ekonomicznie!
(tzn. że jest mała)

Ale czasami interesuje nas tylko kierunek. Innym razem to która zmienna „objaśnia” lepiej.

O istotności wnioskujemy łatwo patrząc na p-value

- P-value opisuje prawdopodobieństwo że nasze oszacowania by się pojawiły przy prawdziwej hipotezie 0.

- Im p-value mniejsze, tym bardziej jesteśmy przekonani, że Hipoteza 0 nie jest prawdziwa (i że możemy potwierdzić alternatywę).

- Zwyczajowo patrzy się na poziomy 10%, 5%, 1%. Choć w ostatnich czasach bardziej pewnie 5%, 1%, 0.1%.

O precyzji łatwo wnioskujemy patrząc na przedziały ufności (Confidence Interval)

- Przy założeniu kryterium 5%, przedziały mówią:
  „95% szacowanych przedziałów będzie zawierała prawdziwy efekt”
  „wartości `kompatybilne` z danymi i modelem są w tym przedziale”

- Jeśli przedział obejmuje 0, to nie możemy na poziomie 5% odrzucić hipotezy zerowej.

O precyzji łatwo wnioskujemy patrząc na przedziały ufności (Confidence Interval)

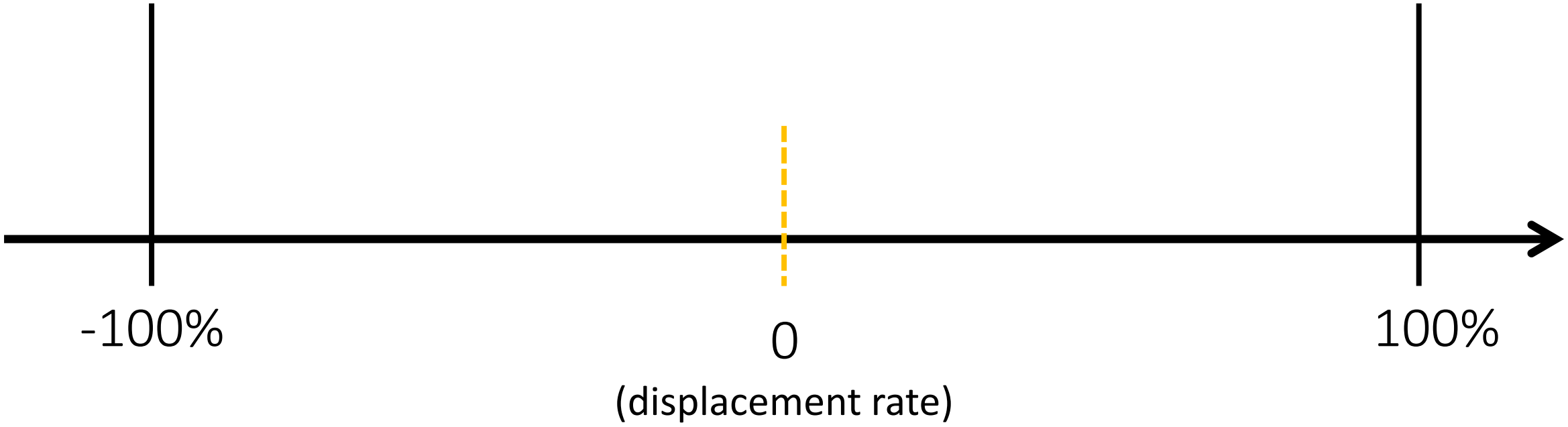
- Przy założeniu kryterium 5%, przedziały mówią:
  „szacujemy, że na 95% prawdziwa wartość jest w tym przedziale”.

- Jeśli przedział obejmuje 0, to nie możemy na poziomie 5% odrzucić hipotezy zerowej.

- To nie oznacza że związek między zmiennymi nie istnieje ! ! ! ! !

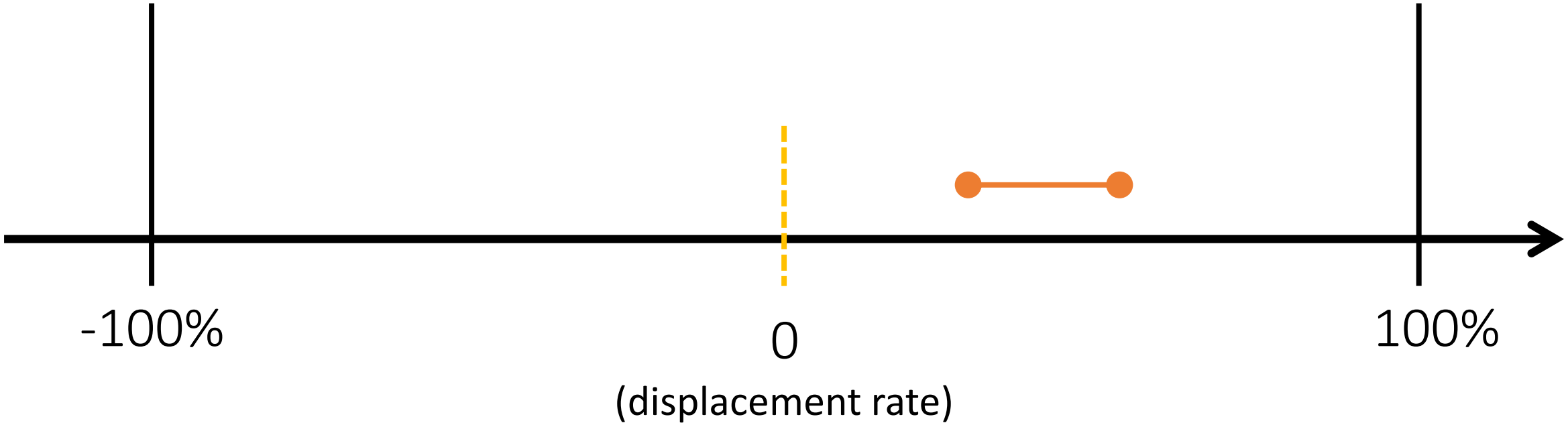
  W praktyce nie możemy potwierdzić hipotezy zerowej – jedynie ją odrzucić!

Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji



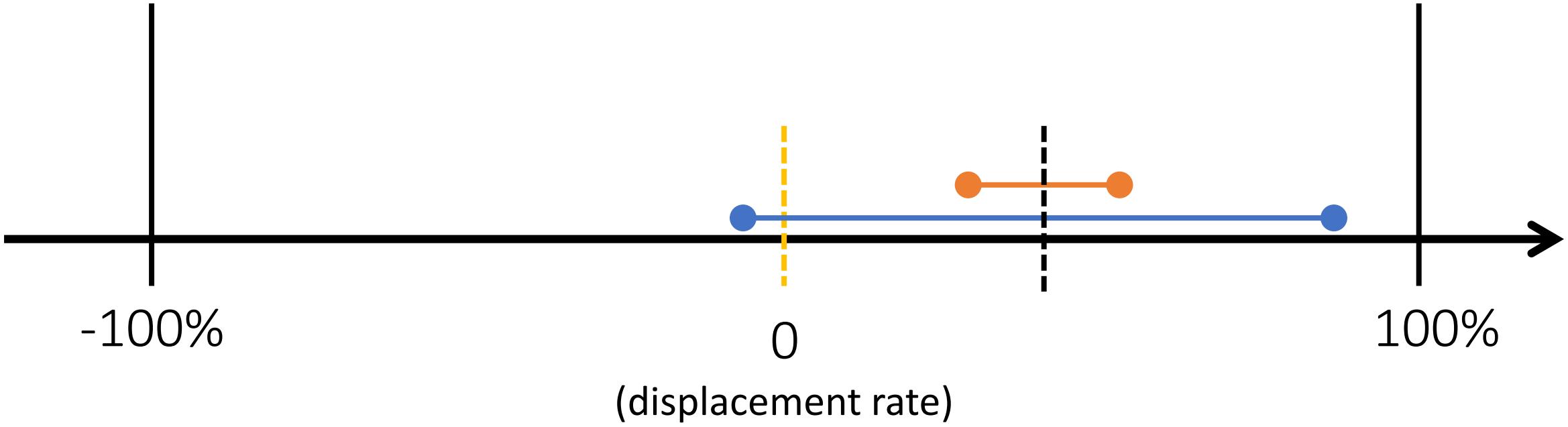
Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji

Mniejsza zmienność / Więcej obserwacji -> Węższy przedział (wyższa precyzja)

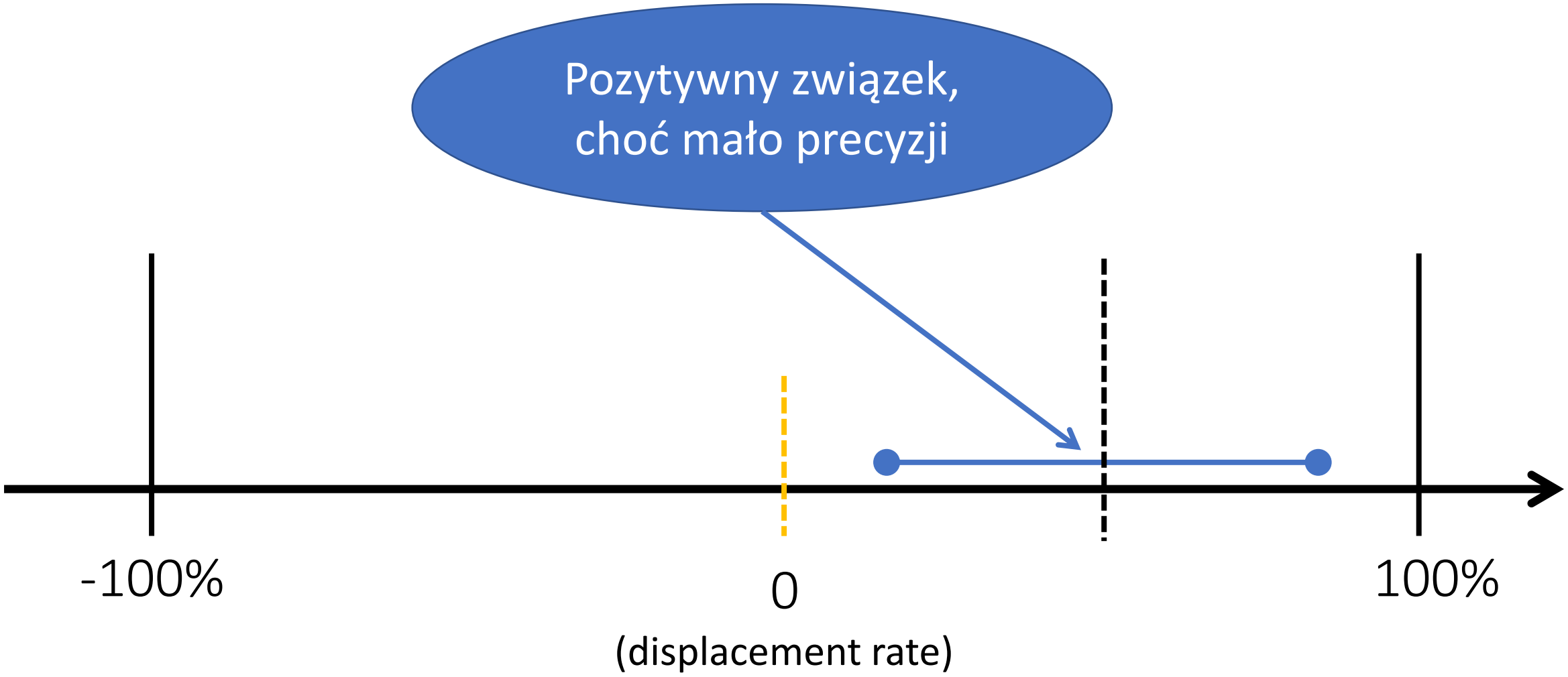


Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji

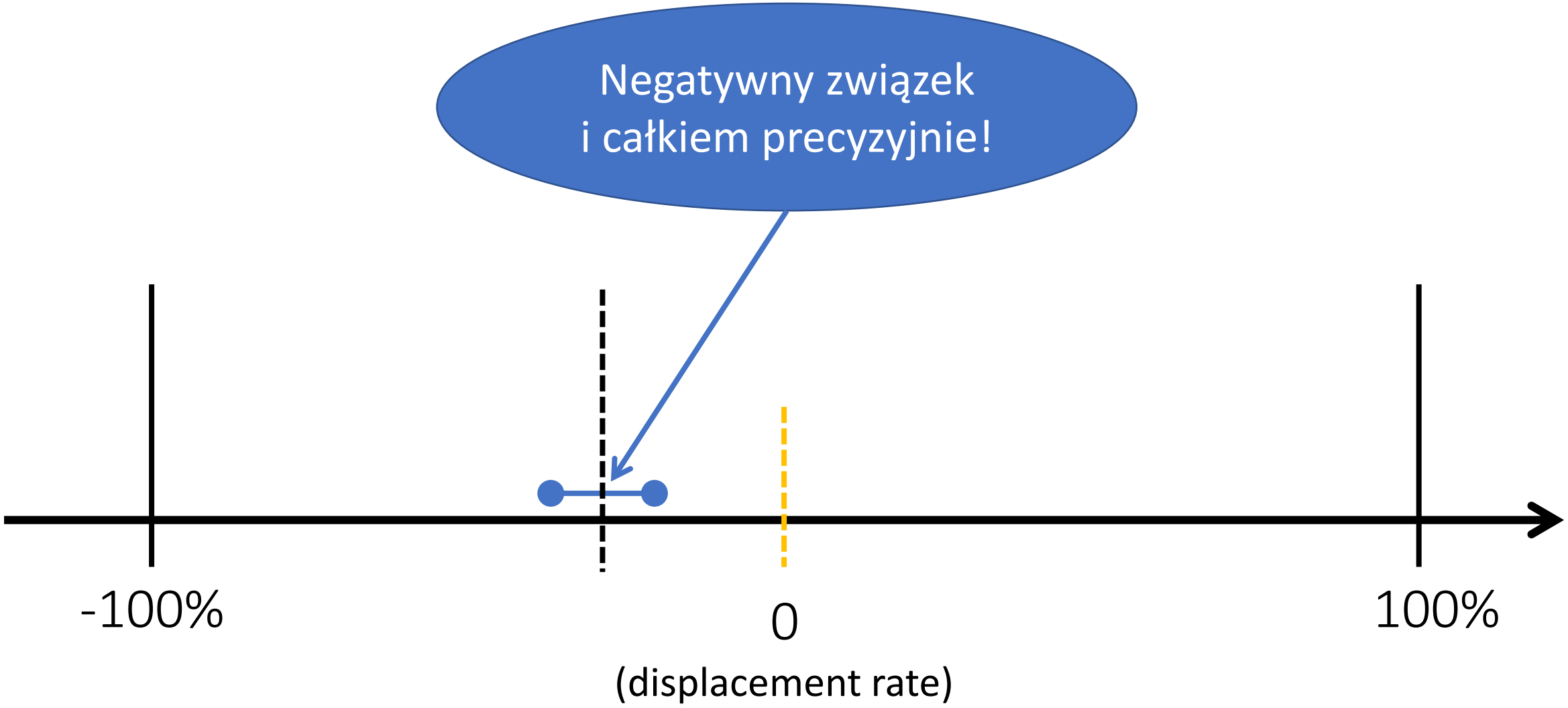
Mniejsza zmienność / Więcej obserwacji -> Węższy przedział (wyższa precyzja)
Wyższa zmienność / Mniej obserwacji -> Szerszy przedział (niższa precyzja)



Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji

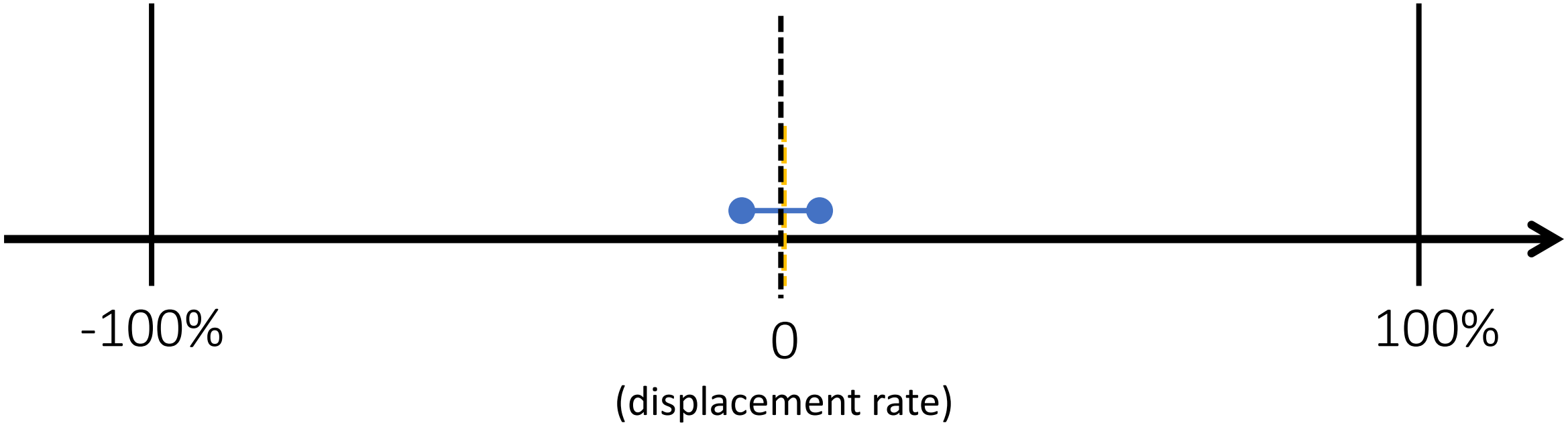


Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji



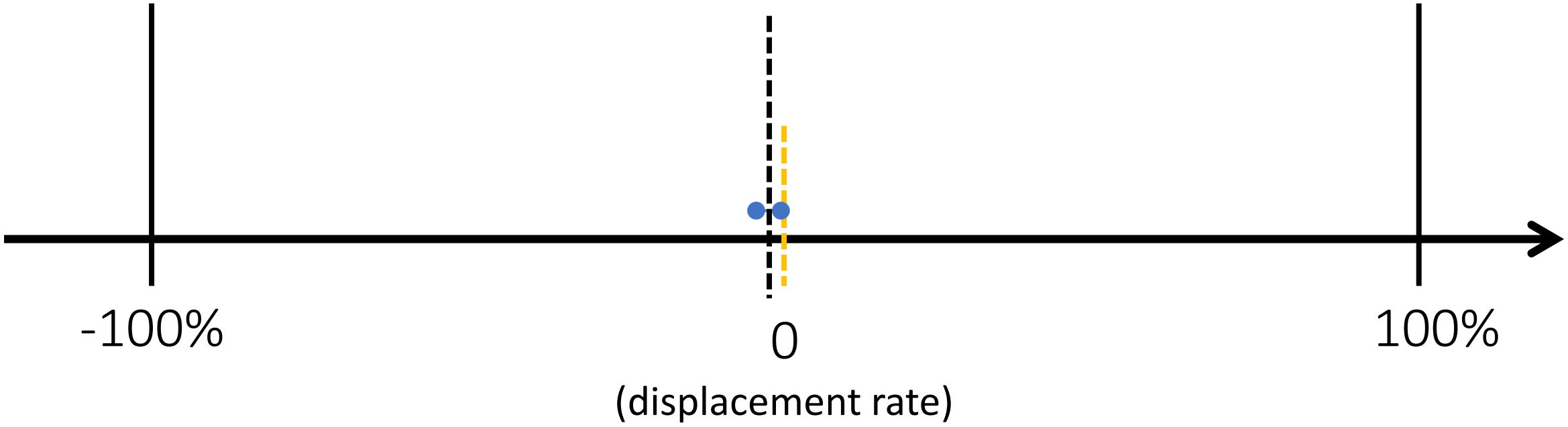
Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji

Na 95% jest bardzo bliskie zera, to prawie jak 0!



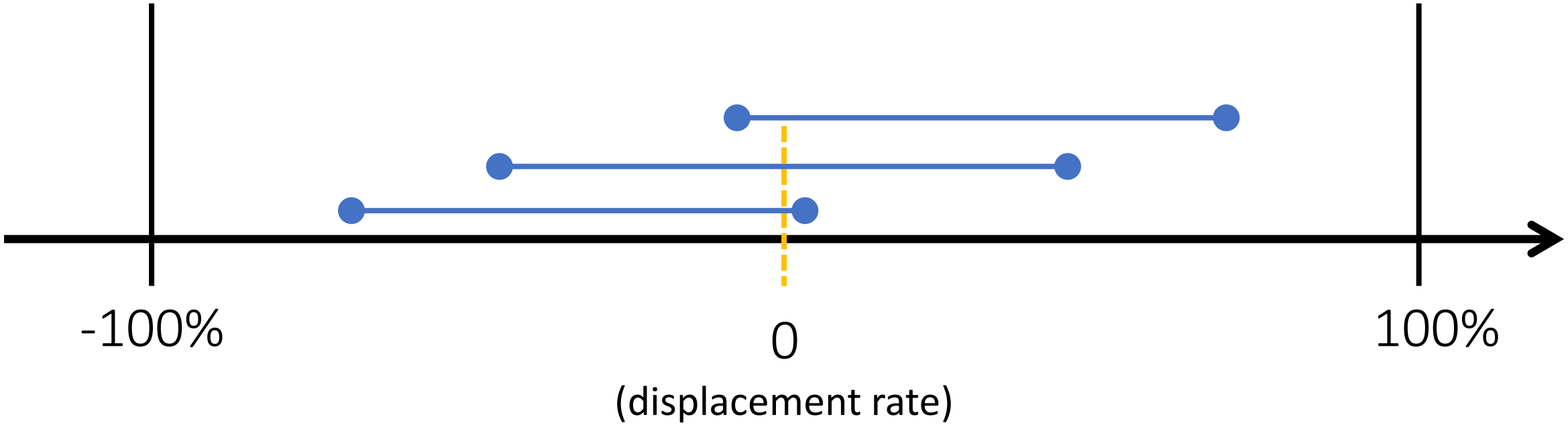
Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji

Tu najpewniej mniejsza od 0, ale tak mała że to w zasadzie bez znaczenia



Dlaczego to ważne?
Wyobraźmy sobie że szacujemy miarę substytucji

Wreszcie: te przedziały są tak szerokie, że nie mówią nam prawie nic!



To ważne!

- Żeby dobrze interpretować wyniki
- Bo w polityce / mediach pojawiają się błędne zrównania braku udowodnionego związku z udowodnieniem braku związku.
  - Np. jak WHO ostrożnie komunikowało „obecnie brak dowodów na …” w kwestii koronawirusa, media podawały to czasem jako „WHO mówi że nie ma związku …”
- Inny przykład związany z piractwem i z tłumaczeniem przedziałów ☺

https://www.youtube.com/watch?v=ZkNv_WEMZC0

O czym jeszcze warto pamiętać!

Hipotezy

- Powinny być falsyfikowalne! Tj. jest dość jasne kiedy hipoteza może zostać uznana za potwierdzoną.

- Możemy odrzucić hipotezę zerową, ale nie możemy jej potwierdzić! (czasem możemy napisać że związek jest pomijalny, jeśli przedziały ufności są bardzo bliskie zera – patrz poprzednie slajdy)

- Nie łączymy hipotez – jeśli tylko część się potwierdzi to nie wiadomo co potem z taką hipotezą zrobić.

Zależność to nie przyczynowość!

- Określić czy:
  a) X wpływa na Y;
  b) Y wpływa na X;
  c) Lub jednak to w ogóle Z wpływa naraz na X i Y

  nie jest proste.

- W hipotezach ostrożniej mówić o pozytywnych/negatywnych zależnościach (nie o przyczynowości, chyba że jesteśmy jej pewni).

Np.:

- Zdrowsze osoby zarabiają więcej. Ale czy to dzięki zdrowiu tyle zarabiają, czy raczej są zdrowe dzięki temu że mają na to pieniądze?

Uwaga na zmienne pominięte!

Np.:

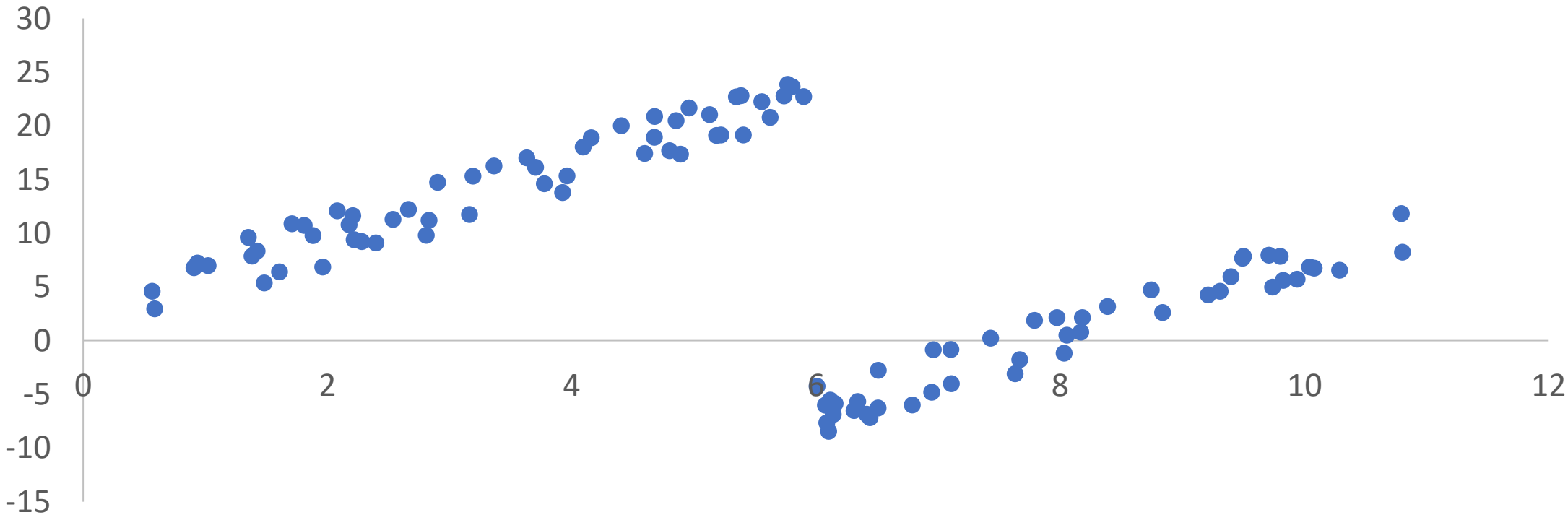
- tytuły częściej piracone sprzedają się lepiej
- osoby częściej piracące kupują więcej treści legalnie

Obie zależności z grubsza znikają jeśli uwzględnimy odpowiednio:

- jakość/popularność tytułów
- indywidualne zamiłowanie do kultury

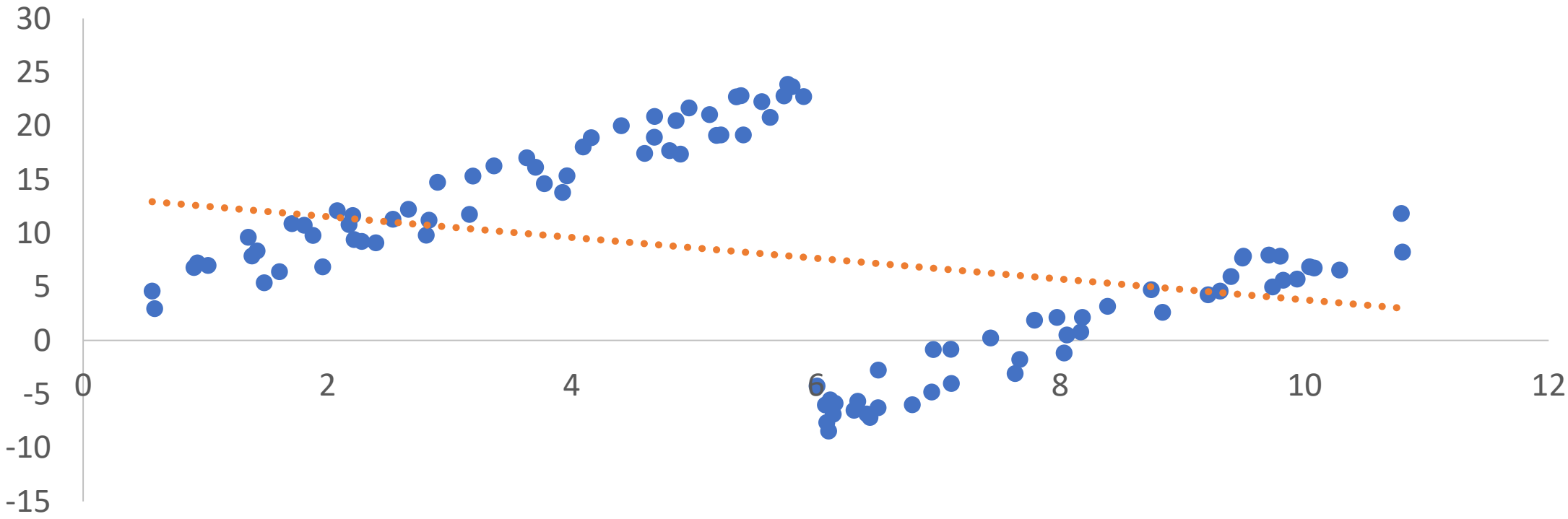
Wykresy bywają pomocne w zorientowaniu się co nam siedzi w danych i co napędza wyniki

Tu np. mamy problem jeśli nie uwzględnimy tego czym różnią się te dwie grupy obserwacji



Wykresy bywają pomocne w zorientowaniu się co nam siedzi w danych i co napędza wyniki

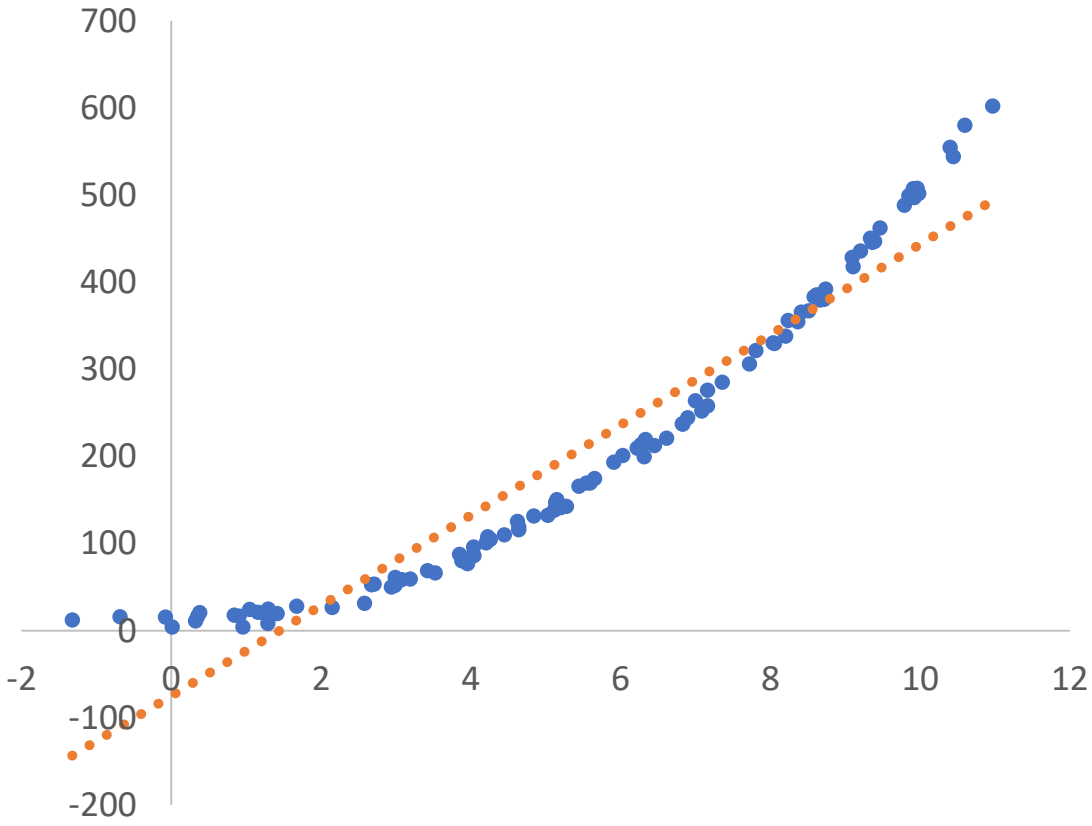
Tu np. mamy problem jeśli nie uwzględnimy tego czym różnią się te dwie grupy obserwacji



Liniowość vs inne zależności

Czasami prawdziwa zależność nie jest liniowa!

To może być dla nas ważne, ale nie musi.

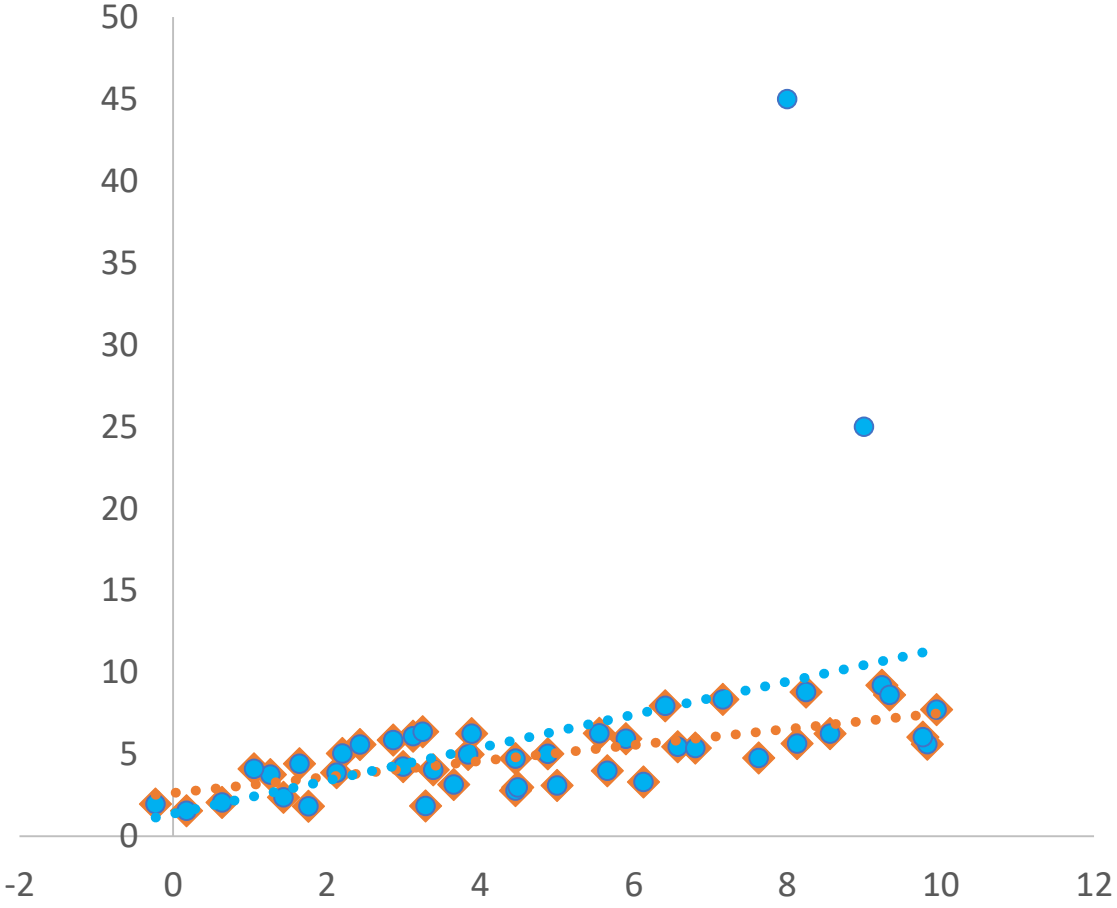


Outliery, czyli obserwacje odstające

Dobrze jest sprawdzić, czy w danych nie siedzi nic `dziwnego’

Ekonometria ma sposoby by nie brać takich obserwacji `na poważnie`

Ale czasem może się okazać że zależności napędza nam podgrupa nietypowych obserwacji



Przykładowy wydruk po regresji (tu ze Staty)

Source	SS	df	MS	Number of obs	=	74
				F(3, 70)	=	23.29
Model	317252881	3	105750960	Prob > F	=	0.0000
Residual	317812515	70	4540178.78	R-squared	=	0.4996
				Adj R-squared	=	0.4781
Total	635065396	73	8699525.97	Root MSE	=	2130.8

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	3.464706	.630749	5.49	0.000	2.206717	4.722695
mpg	21.8536	74.22114	0.29	0.769	-126.1758	169.883
foreign	3673.06	683.9783	5.37	0.000	2308.909	5037.212
_cons	-5853.696	3376.987	-1.73	0.087	-12588.88	881.4934

Oszacowanie zależności

Przykładowy wydruk po regresji (tu ze Staty)

Source	SS	df	MS	Number of obs	=	74
				F(3, 70)	=	23.29
Model	317252881	3	105750960	Prob > F	=	0.0000
Residual	317812515	70	4540178.78	R-squared	=	0.4996
				Adj R-squared	=	0.4781
Total	635065396	73	8699525.97	Root MSE	=	2130.8

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	3.464706	.630749	5.49	0.000	2.206717	4.722695
mpg	21.8536	74.22114	0.29	0.769	-126.1758	169.883
foreign	3673.06	683.9783	5.37	0.000	2308.909	5037.212
_cons	-5853.696	3376.987	-1.73	0.087	-12588.88	881.4934

Oszacowanie zależności

P-value

Przykładowy wydruk po regresji (tu ze Staty)

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =     23.29
       Model |  317252881          3  105750960    Prob > F        =    0.0000
    Residual |  317812515         70  4540178.78   R-squared       =    0.4996
-------------+----------------------------------   Adj R-squared   =    0.4781
       Total |  635065396         73  8699525.97   Root MSE        =    2130.8

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.464706    .630749     5.49   0.000     2.206717    4.722695
         mpg |    21.8536   74.22114     0.29   0.769    -126.1758     169.883
     foreign |    3673.06   683.9783     5.37   0.000     2308.909    5037.212
       _cons |  -5853.696   3376.987    -1.73   0.087    -12588.88    881.4934
------------------------------------------------------------------------------
```

Oszacowanie zależności

P-value

Przedziały ufności