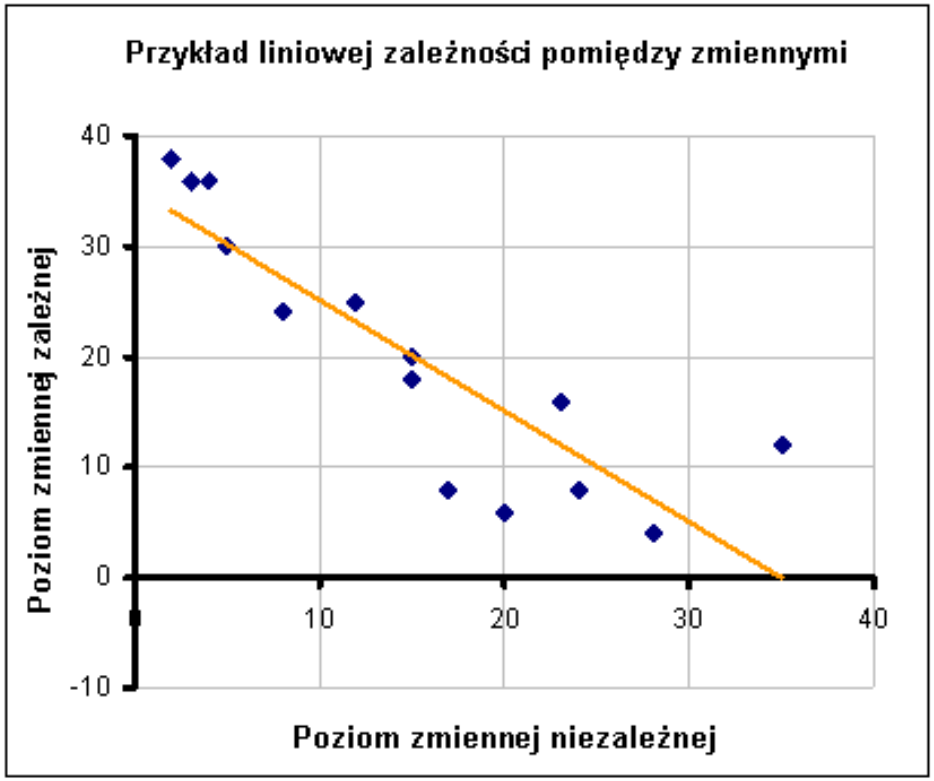
MODEL REGRESJI LINIOWEJ
MIARY DOPASOWANIA FUNKCJI DO DANYCH RZECZYWISTYCH

Funkcja regresji - to narzędzie do badania powiązań między zmiennymi.

Dużym problemem jest wybór postaci analitycznej funkcji dla analizowanego zagadnienia. Ułatwieniem może być sporządzenie m.in. **wykresu rozrzutu**, gdzie dla każdej (i-tej) pary wartości zmiennej niezależnej (X) i zmiennej zależnej (Y) tworzymy punkt o współrzędnych Xi, Yi.

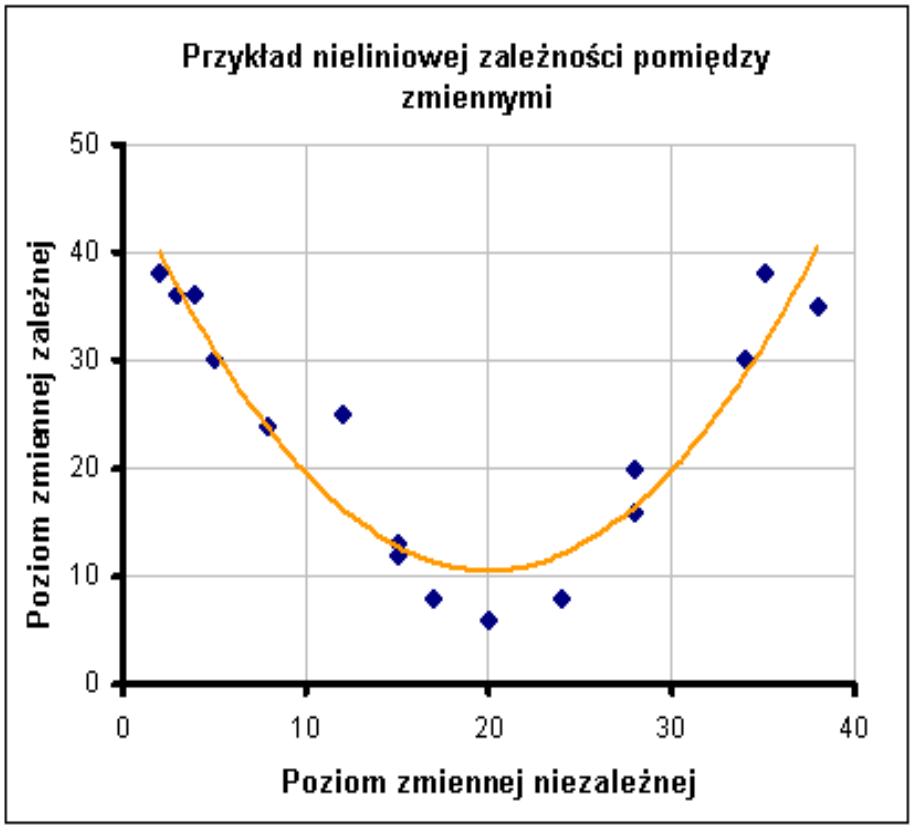
Jeżeli zmiennych niezależnych jest więcej, wówczas konstruujemy odpowiednio większą ilość wykresów rozrzutu, przedstawiających zależność pomiędzy każdą zmienną niezależną (oś odciętych) a zmienną zależną (oś rzędnych). Z wykresu (wykresów) odczytujemy prawdopodobny rodzaj zależności pomiędzy zmiennymi niezależnymi a zmienną zależną.

Jeżeli chmura punktów układa się w przybliżeniu wzdłuż linii prostej (co zostało pokazane na poniższym wykresie), wówczas możemy wykorzystać liniową funkcję regresji.



Źródło: opracowanie własne.

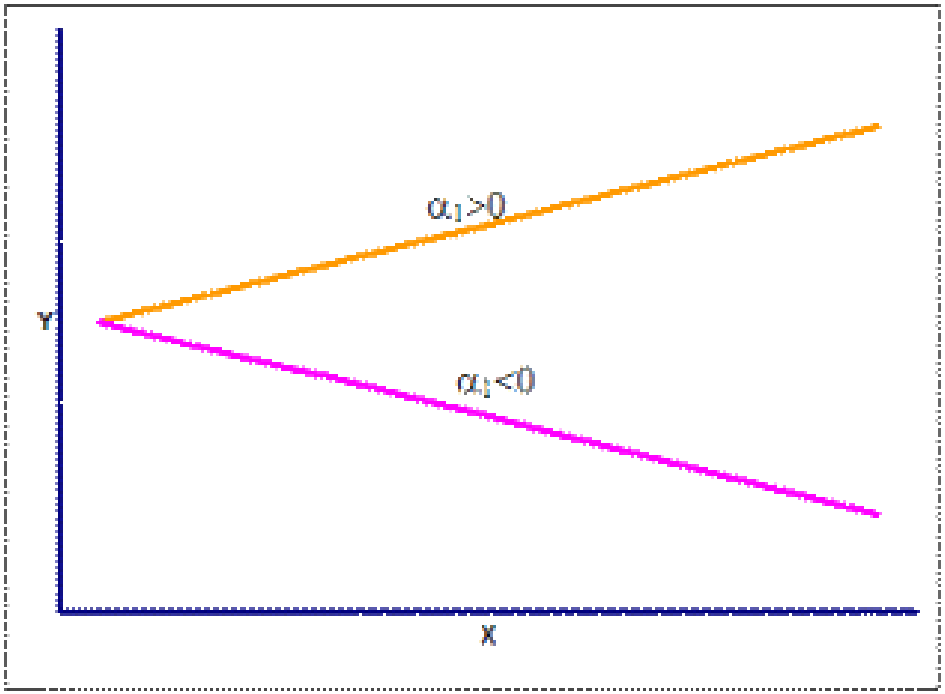
Jeżeli chmura punktów nie układa się wzdłuż linii prostej, wówczas należy wykorzystać inną analityczną postać funkcji regresji (na przykład funkcje potęgowe, logarytmiczne, wielomianowe czy też wykładnicze).



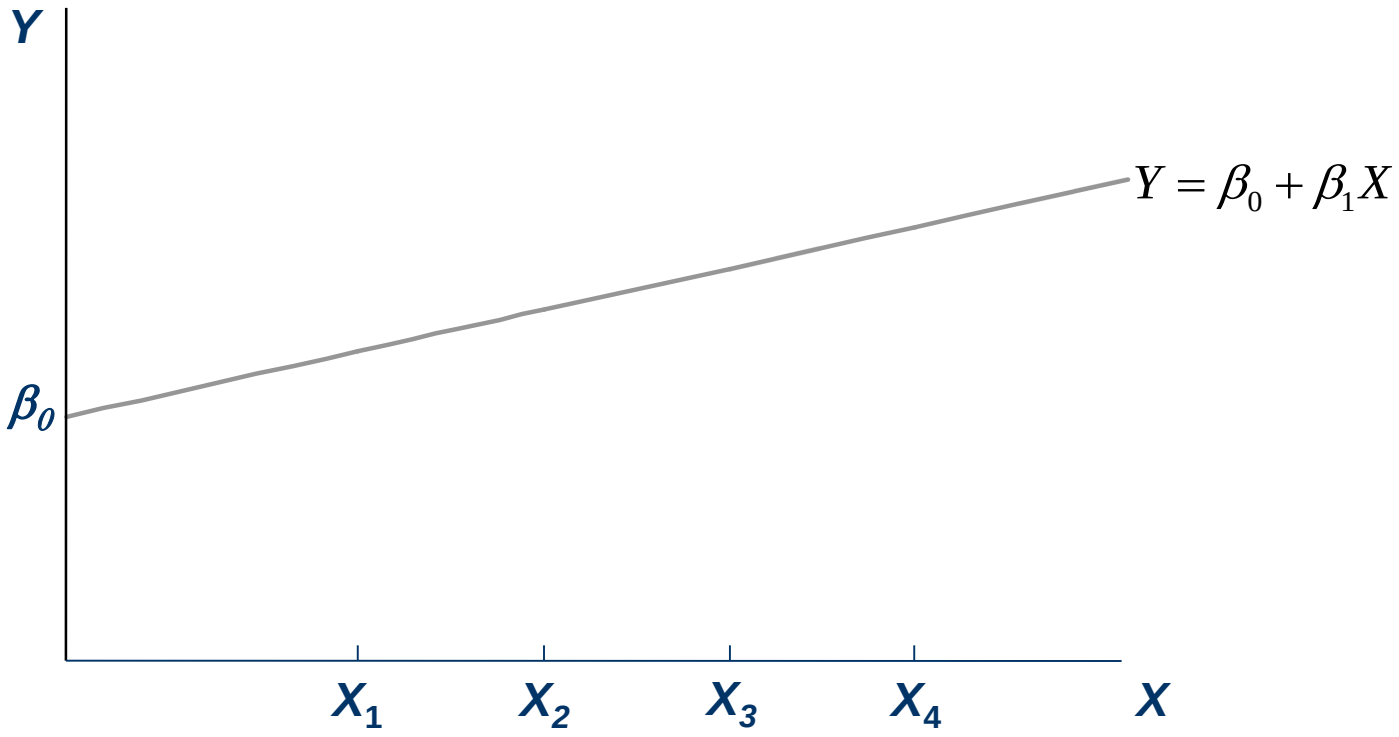
Źródło: opracowanie własne.

Charakter zależności między zmiennymi może przybierać różne formy, od prostych funkcji matematycznych po niezwykle skomplikowane. Najprostsza zależność składa się z relacji prostej lub liniowej (funkcja liniowa).

To jest przykładowy wykres funkcji liniowej :

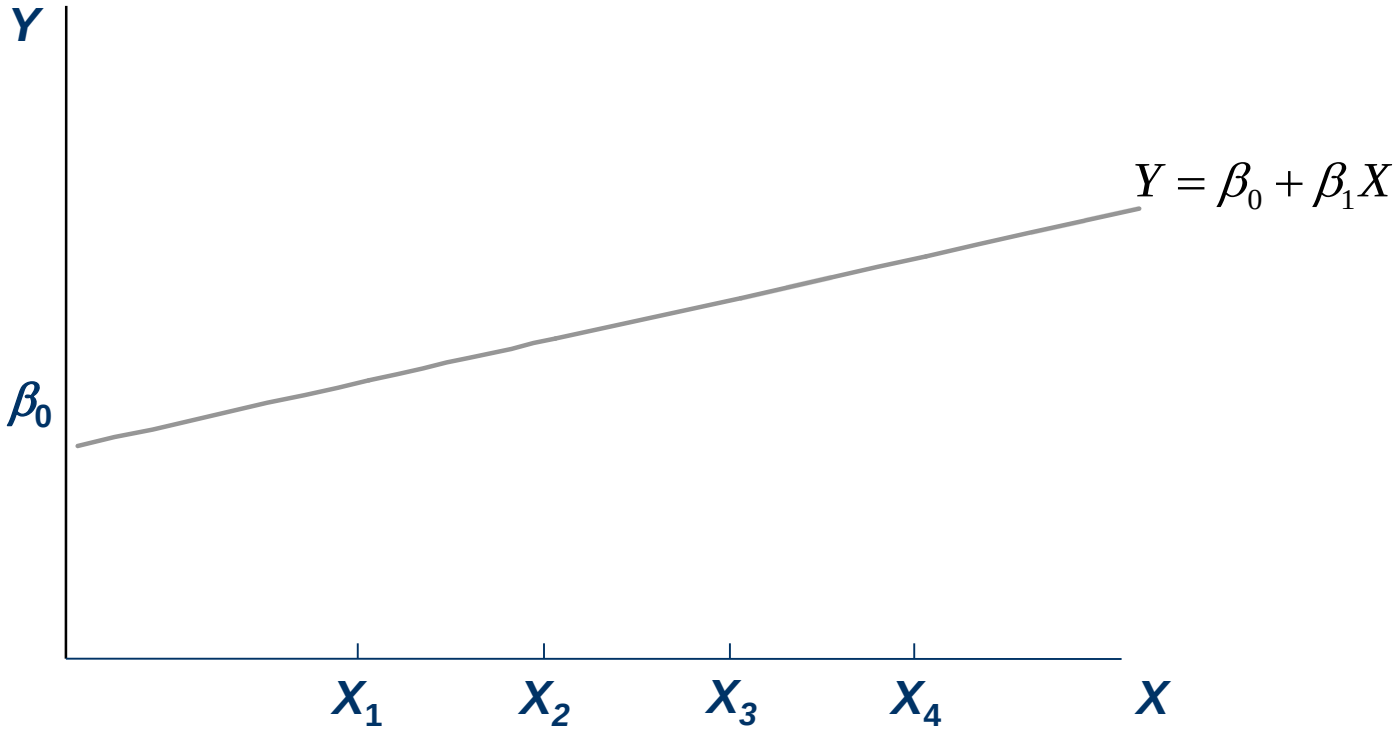


MODEL REGRESJI LINIOWEJ



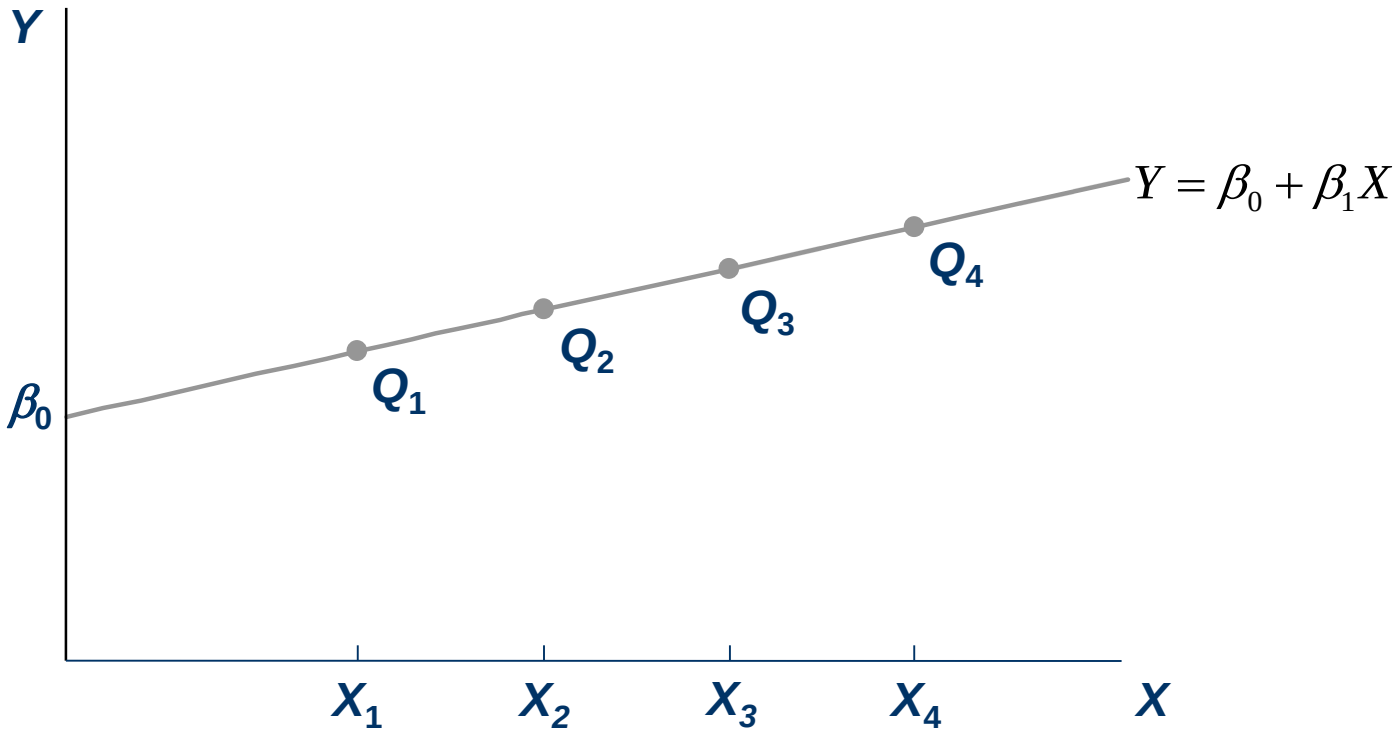
Załóżmy, że zmienna Y jest funkcją liniową innej zmiennej X o nieznanych parametrach β_0 i β_1 , które chcemy oszacować.

MODEL REGRESJI LINIOWEJ



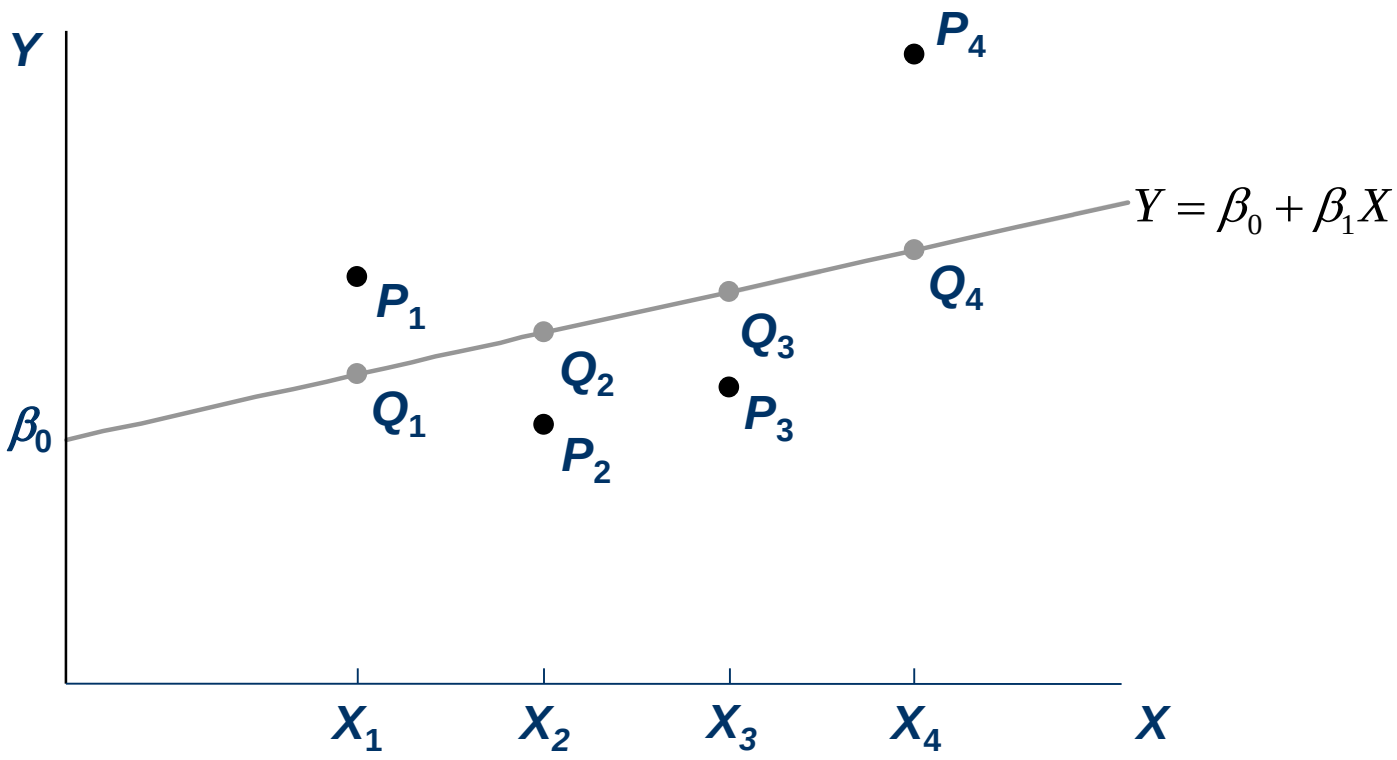
Załóżmy, że mamy próbkę 4 obserwacji z wartościami X, jak pokazano.

MODEL REGRESJI LINIOWEJ



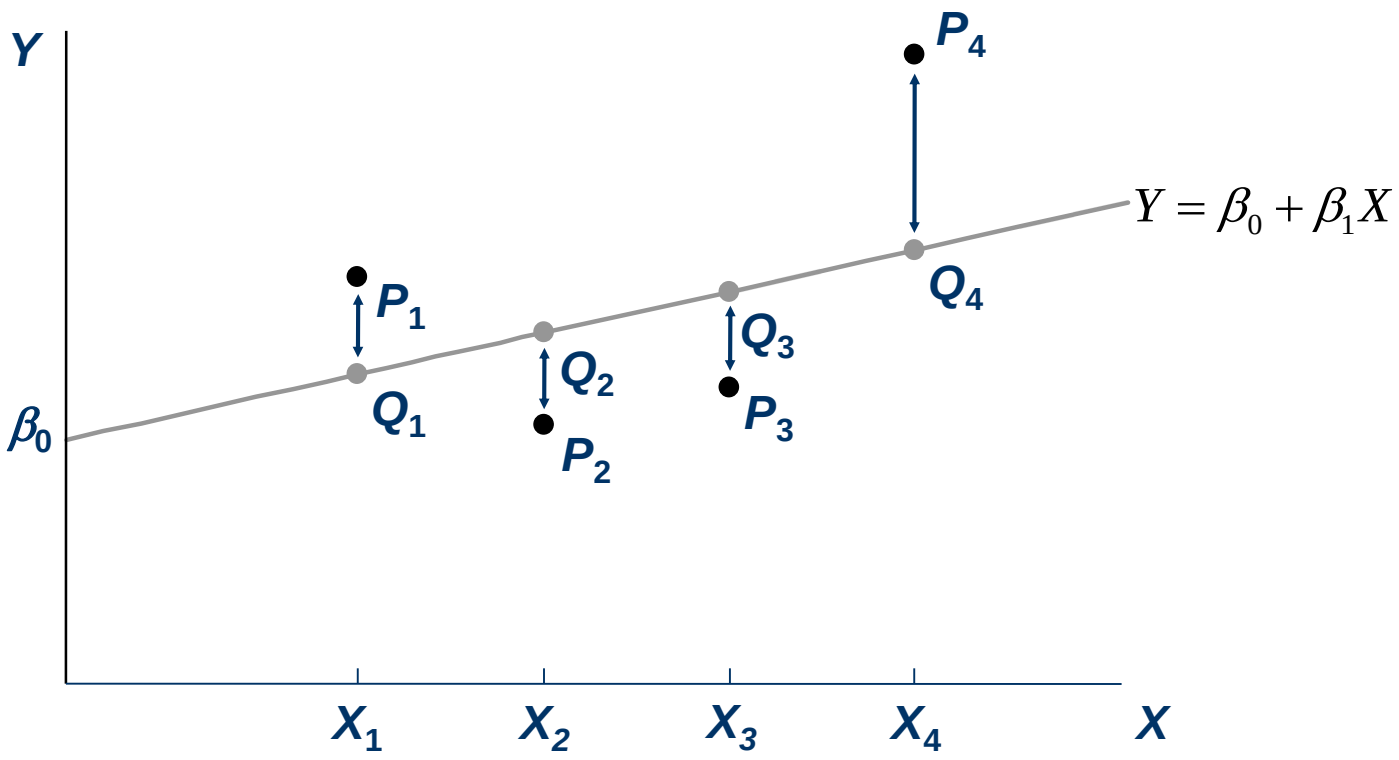
Gdyby związek był dokładny, obserwacje leżałyby na linii prostej i nie mielibyśmy problemów z uzyskaniem dokładnych oszacowań β_0 i β_1 . Gdy wszystkie pary empiryczne punktów X-Y leżą na linii prostej - nazywa się to relacją funkcjonalną lub deterministyczną.

MODEL REGRESJI LINIOWEJ



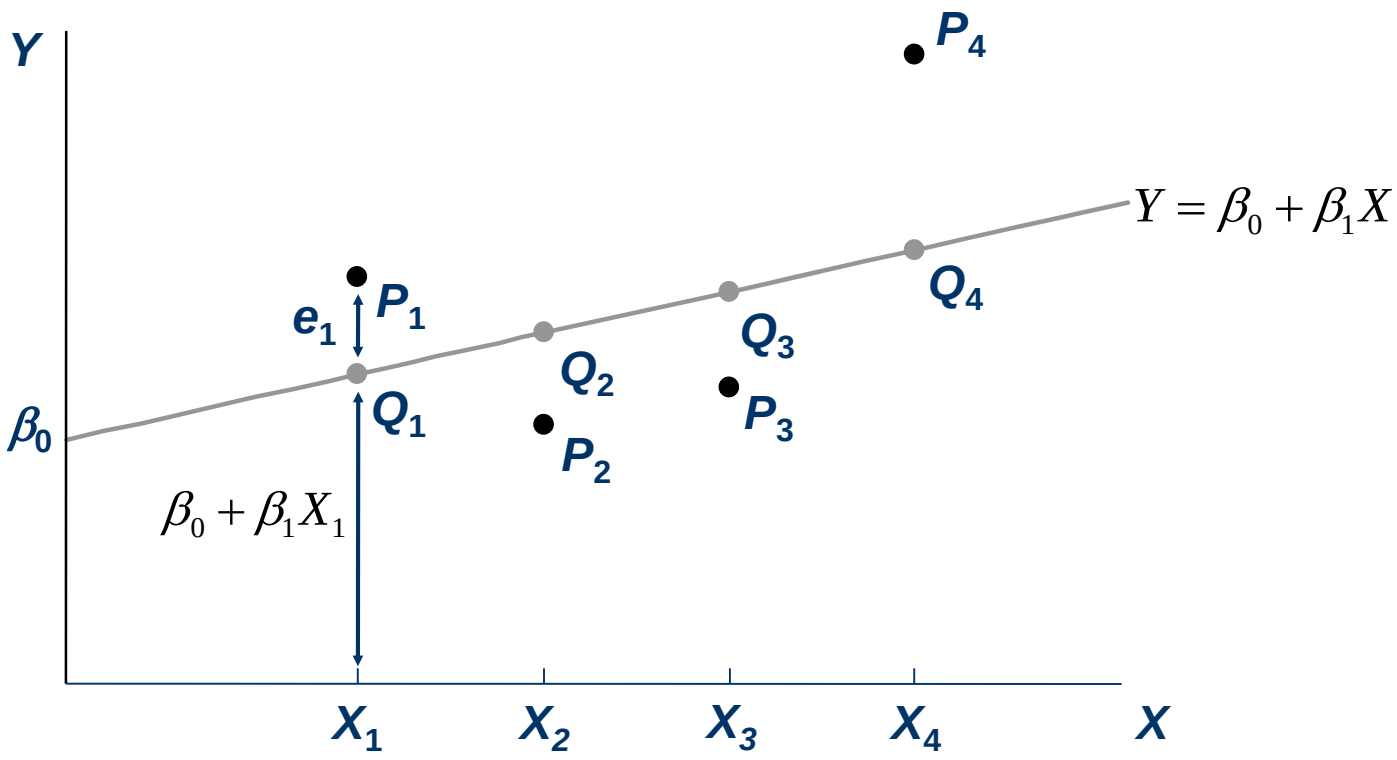
W praktyce większość relacji ekonomicznych nie jest ścisła, a rzeczywiste wartości Y różnią się od tych odpowiadających linii prostej.

MODEL REGRESJI LINIOWEJ



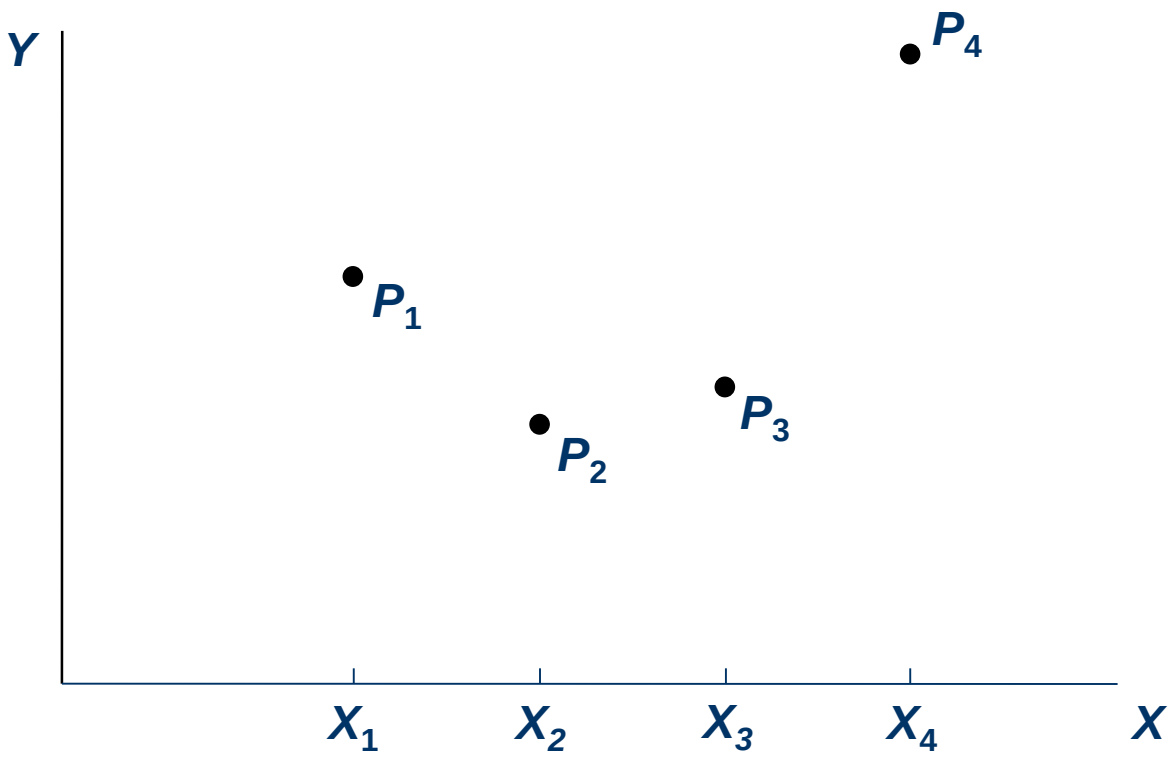
Aby uwzględnić takie rozbieżności, napiszemy model jako $Y = \beta_0 + \beta_1 X + e$, gdzie e to składnik losowy.

MODEL REGRESJI LINIOWEJ



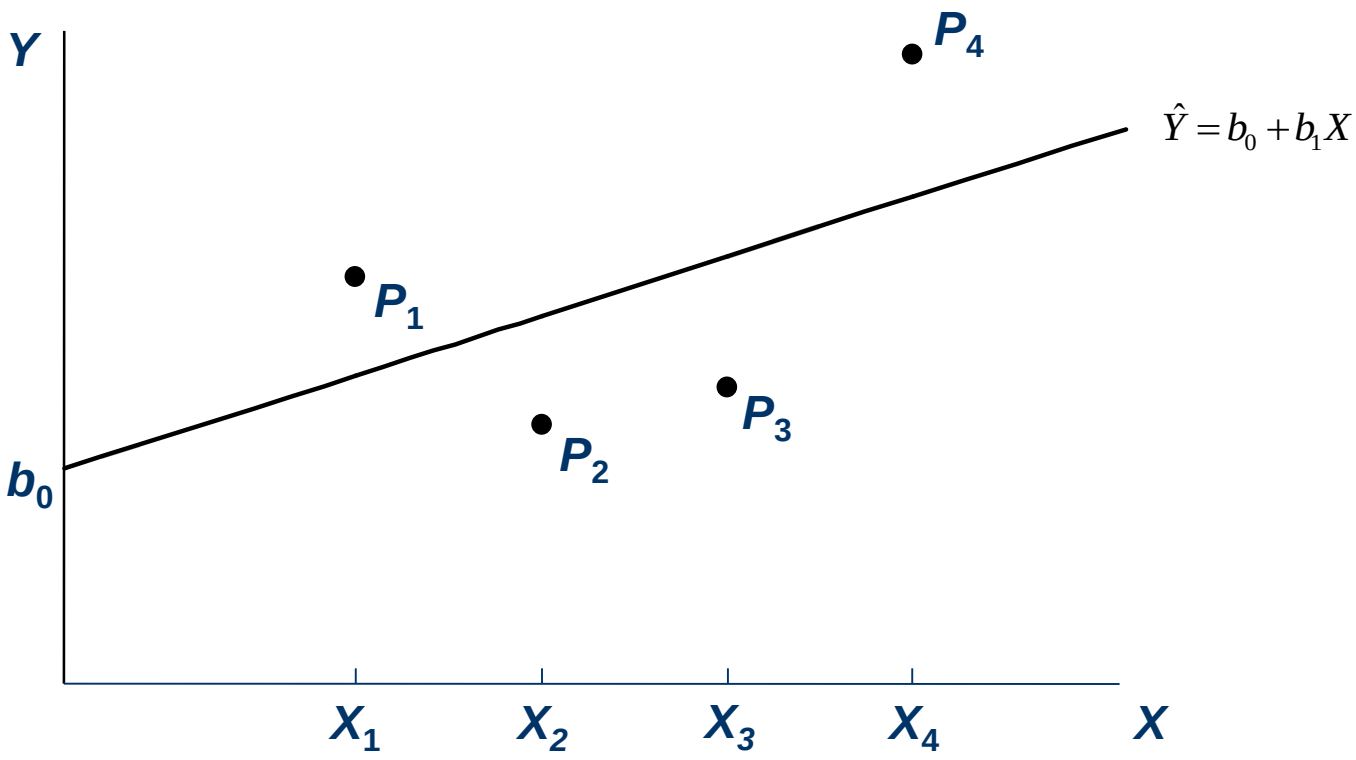
Każda wartość Y ma zatem składową nielosową, $\beta_0 + \beta_1 X$ i składową losową, e. Pierwsza obserwacja została podzielona na te dwa elementy.

MODEL REGRESJI LINIOWEJ



W praktyce widzimy tylko punkty P.

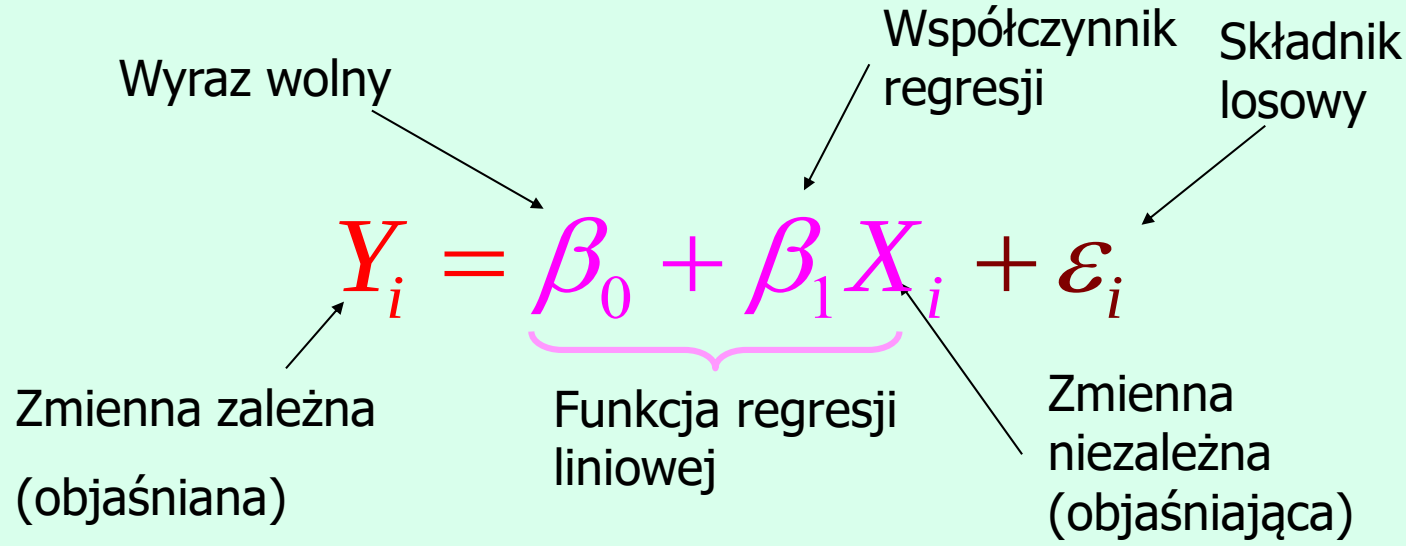
MODEL REGRESJI LINIOWEJ



Oczywiście możemy użyć punktów P, aby narysować linię, która jest przybliżeniem do linii $\hat{Y} = \beta_0 + \beta_1 X$. Jeśli postać oszacowaną modelu zapiszemy w następujący sposób $Y = b_0 + b_1 X$, to oznacza, że b_0 jest estymatorem β_0 a b_1 jest estymatorem β_1.

MODEL REGRESJI LINIOWEJ

Funkcja regresji liniowej jest linią prostą, która opisuje zależność średniej wartości jednej zmiennej od drugiej



MODEL REGRESJI LINIOWEJ – INTERPRETACJA PARAMETRÓW

Jednak uzyskaliśmy dane tylko z losowej próby, a nie z całej populacji. Dla próbki b_0 i b_1 można zastosować jako estymatory odpowiednich parametrów populacji β_0 i β_1

$$\hat{y}_i = b_0 + b_1 x_i + e_i$$

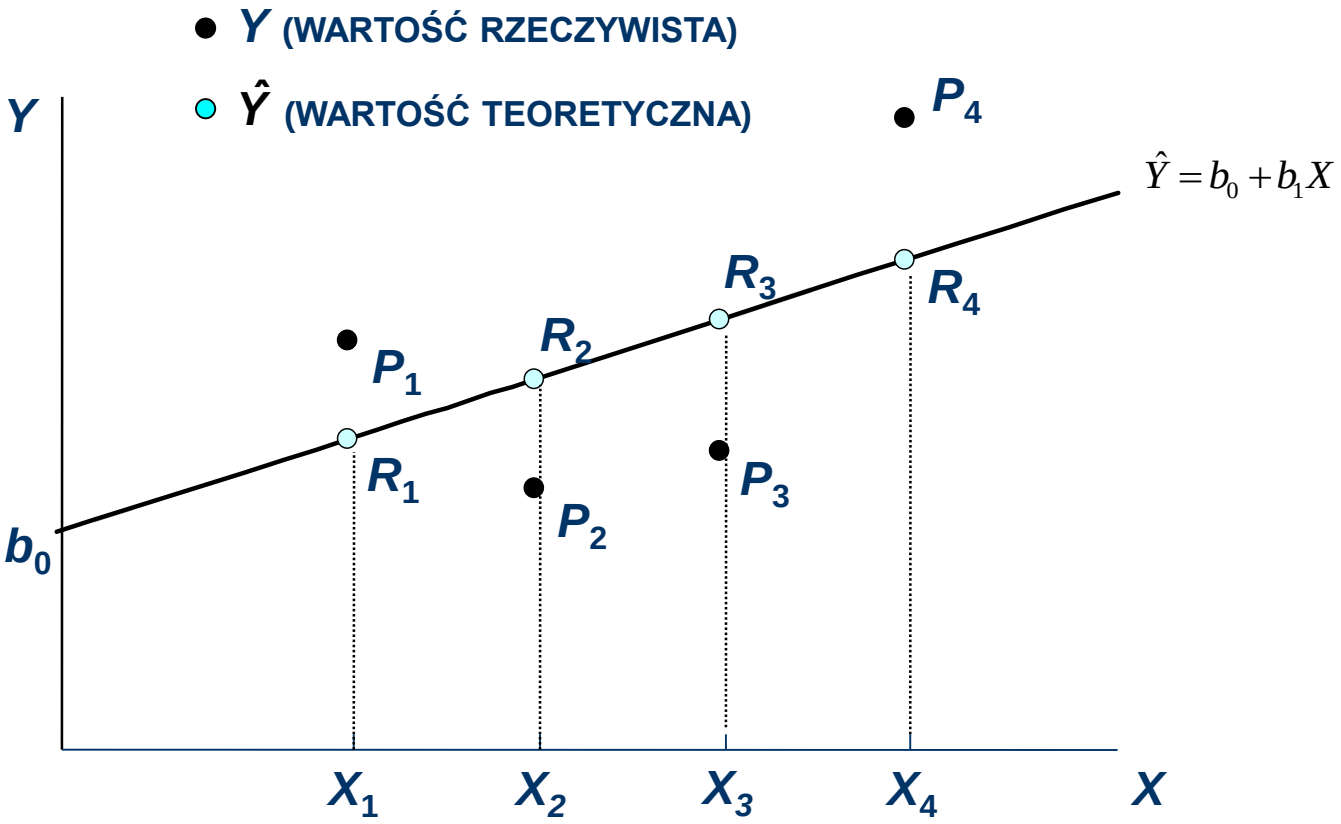
Punkt przecięcia b_0 i nachylenie b_1 są współczynnikami linii regresji. Nachylenie b_1 jest zmianą Y (informuje, że o tyle wzrośnie wartość zmiennej Y, jeśli $b_1>0$ lub zmaleje, jeśli $b_1<0$) związaną ze zmianą zmiennej X o jednostkę.

MODEL REGRESJI LINIOWEJ – INTERPRETACJA PARAMETRÓW

Punkt przecięcia to wartość Y, gdy X = 0; jest to punkt, w którym linia regresji populacji przecina oś Y. W niektórych przypadkach punkt przecięcia nie ma znaczenia w świecie rzeczywistym (na przykład gdy X jest rozmiarem klasy, Y jest wynikiem testu - punkt przecięcia jest przewidywaną wartością wyników testu, gdy nie ma uczniów w klasie!).

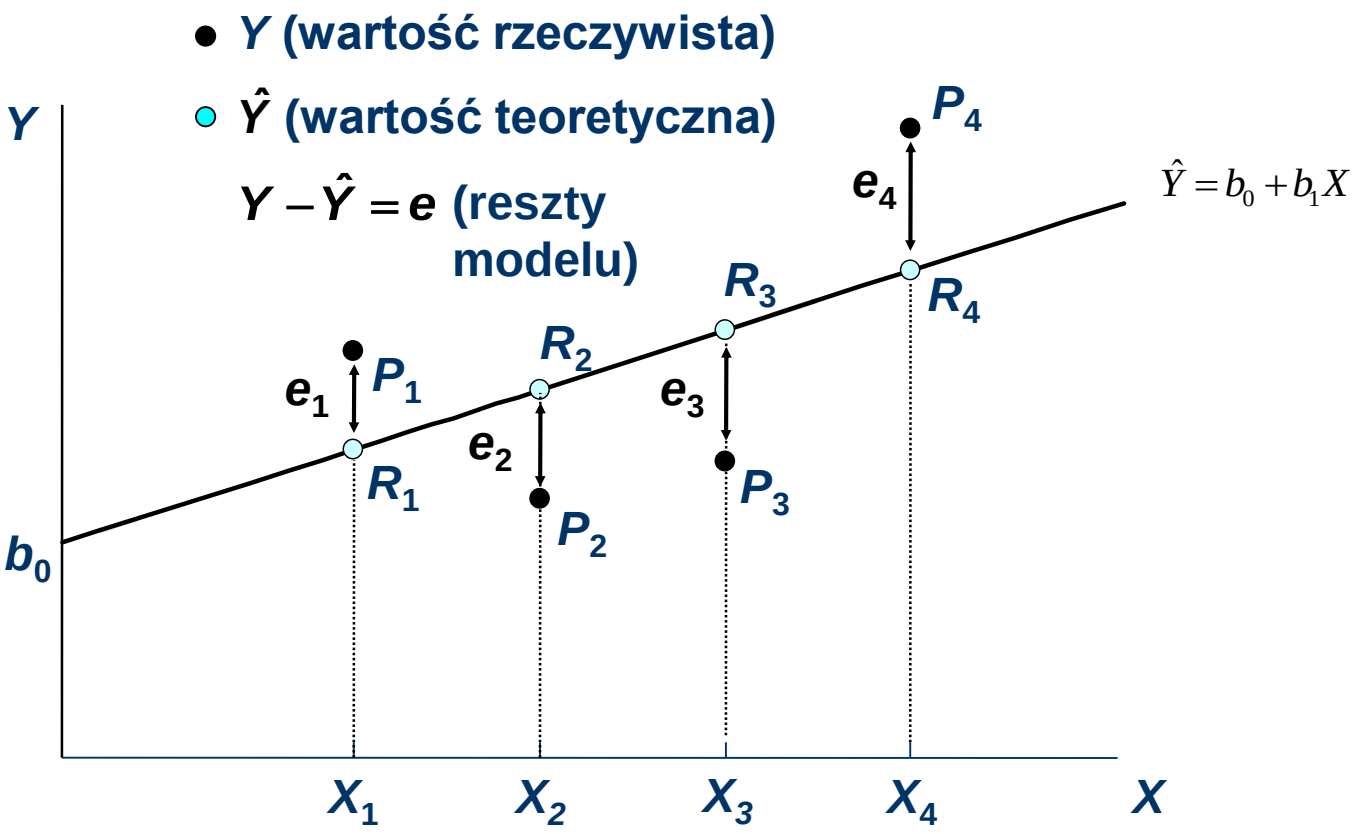
Błąd losowy zawiera wszystkie inne czynniki oprócz X, które określają wartość zmiennej zależnej Y, dla konkretnej obserwacji.

MODEL REGRESJI LINIOWEJ



Linia ta nazywana jest dopasowanym modelem, a przewidywane przez nią wartości Y nazywane są dopasowanymi wartościami Y. Są one podawane na podstawie wysokości punktów R.

MODEL REGRESJI LINIOWEJ



Rozbieżności między wartościami rzeczywistymi i teoretycznymi Y są znane jako wartości resztkowe (RESZTY MODELU).

Kryterium najmniejszych kwadratów:

Zminimalizować *SSE* (residual sum of squares – suma kwadratów reszt), gdzie:

$$SSE = \sum_{i=1}^{n} e_i^2 = e_1^2 + \ldots + e_n^2$$

Na początek narysujemy dopasowaną linię, aby zminimalizować sumę kwadratów reszt, SSE. Jest to określane jako kryterium najmniejszych kwadratów.

Kryterium najmniejszych kwadratów:

Zminimalizować *SSE* (residual sum of squares), gdzie

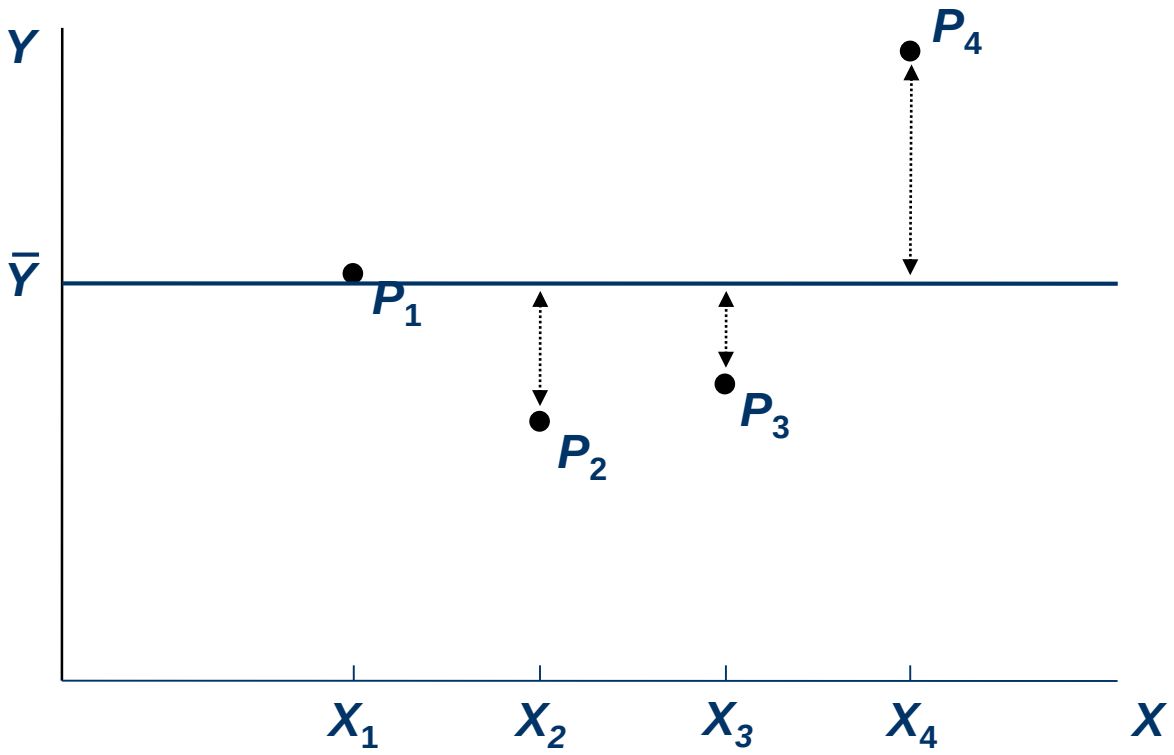
$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = e_1^2 + \ldots + e_n^2$$

Dlaczego nie minimalizować sumy reszt?

$$\sum_{i=1}^{n} e_i = e_1 + \ldots + e_n$$

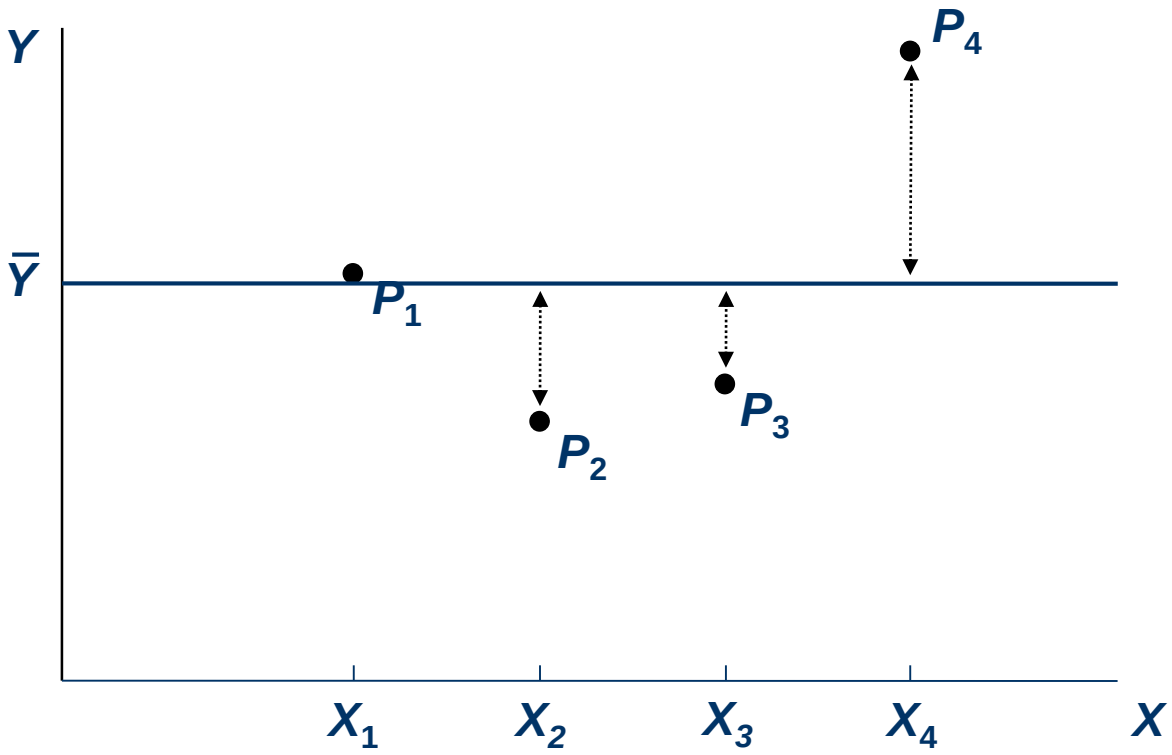
Dlaczego kwadraty reszt? Dlaczego nie minimalizować sumy reszt?

MODEL REGRESJI LINIOWEJ



Odpowiedź jest taka, że pozornie idealne dopasowanie można uzyskać, rysując linię poziomą przez średnią wartość Y. Suma reszt będzie wynosić zero.

MODEL REGRESJI LINIOWEJ



Należy uniemożliwić anulowanie reszt dodatnich przez reszty ujemne, a jednym ze sposobów jest użycie kwadratów reszt.

MODEL REGRESJI LINIOWEJ

W zapisie macierzowym MNK (METODA NAJMNIEJSZYCH KWADRATÓW) może zostać zapisana jako: Y = Xb + e

Mnożąc obie strony równania przez $\mathbf{X^T}$ otrzymujemy:

$$\mathbf{X^T Y = X^T X b}$$

A kiedy rozwiązujemy równanie dla b, otrzymujemy:

$$\mathbf{b = (X^T X)^{-1} X^T Y}$$

gdzie Y jest wektorem kolumnowym wartości Y, a X jest macierzą zawierającą kolumnę jedności, po której następuje kolumna wartości zmiennej X, a b jest wektorem zawierającym estymatory parametrów regresji:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

MODEL REGRESJI LINIOWEJ

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \qquad X^T Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{bmatrix}$$

Jak odwrócić X^TX?

1. **wyznacznik**

$$deX^T X = n \cdot \sum x^2 - (\sum x)^2$$

2. **macierz minorów**

$$\min X^T X = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = \begin{bmatrix} \sum x^2 & \sum x \\ \sum x & n \end{bmatrix}$$

3. **macierz dopełnień algebraicznych**

$$(X^T X)D = \begin{bmatrix} \sum x^2 \cdot (-1)^{1+1} & \sum x \cdot (-1)^{1+2} \\ \sum x \cdot (-1)^{2+1} & n \cdot (-1)^{2+2} \end{bmatrix}$$

4. **macierz odwrotna**

$$(X^T X)^{-1} = \frac{1}{\det X^T X} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix}$$

Po wyznaczeniu parametrów funkcji regresji liniowej należy ocenić **poziom dopasowania** funkcji regresji do rzeczywistych danych. Sprowadza się to do odniesienia generowanych przez funkcję regresji **wartości teoretycznych** do **wartości zaobserwowanych**. Wykorzystuje się w tym celu szereg miar, do najczęściej stosowanych należą: odchylenie standardowe reszt, współczynnik zbieżności oraz współczynnik determinacji.

Wartości teoretyczne obliczamy podstawiając do funkcji regresji liniowej wartości zmiennej niezależnej X.

Odchylenie standardowe reszt (s_ε) – informuje nas, o ile średnio rzecz biorąc wartości cechy Y odchylają się od jej wartości obliczonych na podstawie funkcji regresji.

$$s_\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Gdzie:

$\hat{y}_i$ – wartości teoretyczne obliczane z wykorzystaniem funkcji regresji.

Pozostałe symbole jak wcześniej

Współczynnik zbieżności (φ^2) – wskazuje jaka część zmienności w wartościach cechy Y, nie jest związana ze zmiennością w wartościach cechy X w sensie funkcji regresji.

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

Współczynnik determinacji (R^2) – wskazuje jaka część zmienności cechy Y związana jest ze zmiennością cechy X w sensie przyjętej funkcji regresji.
Pomiędzy współczynnikiem determinacji i zbieżności ma miejsce następująca zależność:

$$R^2 = 1 - \varphi^2$$

Współczynnik zbieżności oraz determinacji są wielkościami niemianowanymi, dlatego można je wykorzystać do porównywania oceny dopasowania modelu regresji liniowej dla różnych zbiorowości statystycznych.