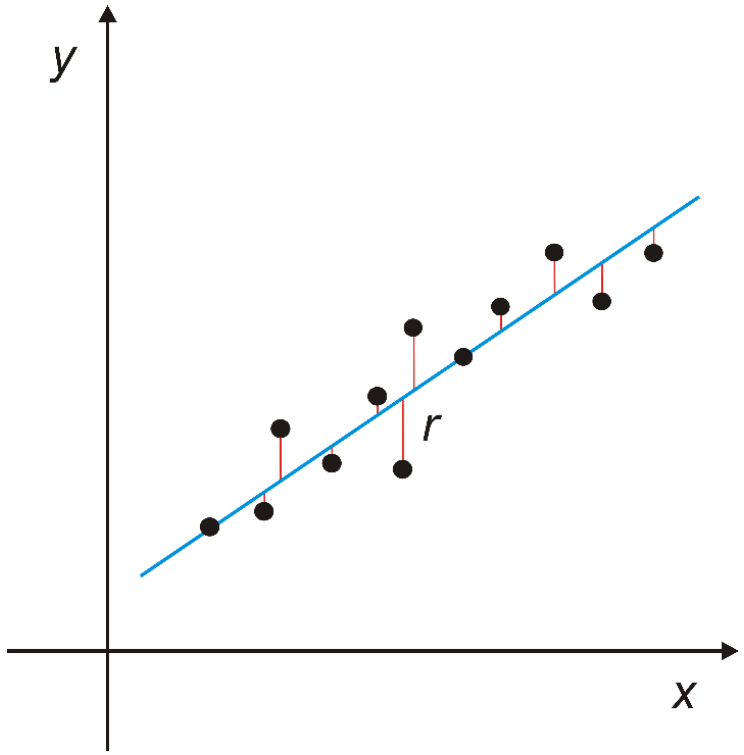
Metoda najmniejszych kwadratów

Regresja (łac. *regressus* - odejście, odstępstwo) - metoda statystyczna zajmująca się badaniem związków i zależności, które występują pomiędzy zmiennymi losowymi np. Y i X. Regresja dzieli się na **regresję liniową** oraz **regresją nieliniową**. Wykresem regresji liniowej dwóch zmiennych jest prosta o równaniu: $y(x) = ax + b$. W przypadku regresji nieliniowej dwóch zmiennych wykresem jest zazwyczaj pewien wielomian (np. funkcja kwadratowa).



Rysunek 1. Schematyczne przedstawienie zależności liniowej pomiędzy zmiennymi x i y.

Zmienne oraz stałe występujące w równaniu **prostej regresji** $y(x) = ax + b$ nazywamy

x - zmienna niezależna (**objaśniająca**)

y - zmienna zależna (**objaśniana**)

a - współczynnik kierunkowy prostej regresji

b - wyraz wolny

W regresji liniowej zazwyczaj interesuje nas **suma kwadratów różnic** pomiędzy wartościami eksperymentalnymi (●) a wartościami teoretycznymi (które leżą na prostej regresji) - wtedy obliczenia algebraiczne są najprostsze. Ma to jednak swoją wadę: kwadraty różnic pomiędzy wartościami eksperymentalnymi a wartościami teoretycznymi silnie zależą od obserwacji dla których błąd jest największy. Metoda najmniejszych kwadratów daje więc niedokładne lub fałszywe wyniki, jeśli w analizowanym układzie występują **obserwacje odstające** np. pomyłki przy wprowadzaniu danych. W związku z tym w regresji liniowej stosowane są także inne miary błędów, które są bardziej odporne na obserwacje odstające jak np. wartość bezwzględna (moduł) różnic pomiędzy wartościami eksperymentalnymi a wartościami teoretycznymi.

Reszta r (błąd regresji liniowej) - różnica pomiędzy wartością eksperymentalną y_i a wartością teoretyczną $y(x_i)$, która leży na prostej regresji: $y(x_i) = ax_i + b$.

$$r = y_i - y(x_i)$$

Zazwyczaj mamy do czynienia z dodatnimi i ujemnymi błędami (resztami) r regresji liniowej - część wartości eksperymentalnych leży nad prostą regresji ($r > 0$), a część pod prostą regresji ($r < 0$). Interesuje nas jednak *suma błędów podniesionych do kwadratu* (zawsze nieujemna liczba rzeczywista) - proste sumowanie błędów dodatnich i ujemnych nie podniesionych do kwadratu prowadziłoby do ich wzajemnego znoszenia się na skutek różnicy znaków.

Termin regresja jako pierwszy użył Francis Galton. W 1886 roku badał on związek pomiędzy wzrostem rodziców i ich dzieci. Zauważył, że wysocy rodzice mają zazwyczaj również wysokie dzieci. Jednakże wzrost dzieci ponadprzeciętnie wysokich rodziców jest zbliżony bardziej do średniej, niż do wzrostu ich rodziców. Taką tendencję powrotu do średniej Galton określił jako *regresję do średniej*.

Francis Galton (1822-1911) - brytyjski podróżnik, przyrodnik, antropolog, pisarz, lekarz, statystyk oraz prekursor badań nad inteligencją. Opracował metodę statystyczną zwaną regresją. Jako pierwszy zastosował metody statystyczne do studiowania ludzkich różnic indywidualnych i dziedziczenia inteligencji, prywatnie zięć Karola Darwina.

Na podstawie przeprowadzonych rozważań widzimy, że w przypadku **regresji liniowej** zwanej **metodą najmniejszych kwadratów** interesuje nas suma (gdzie N to całkowita liczba obserwacji)

$$S = [y_1 - y(x_1)]^2 + [y_2 - y(x_2)]^2 + [y_3 - y(x_3)]^2 + \cdots + [y_N - y(x_N)]^2$$

Powyższa suma może być zapisana w zwięzłej postaci

$$S = \sum_{i=1}^{N} [y_i - y(x_i)]^2 = \sum_{i=1} [y_i - y(x_i)]^2$$

Wiemy, że równanie prostej regresji ma ogólną postać: $y(x_i) = ax_i + b$, a więc:

$$S = \sum_{i=1} [y_i - (ax_i + b)]^2 = \sum_{i=1} [y_i - ax_i - b]^2$$

Suma kwadratów błędów S zależy wyłącznie od dwóch parametrów: $S = f(a, b)$, które powinny być odpowiednio dobrane - są to **współczynnik kierunkowy** prostej regresji a oraz **wyraz wolny** b.

$$S = f(a, b) = \sum_{i=1} [y_i - ax_i - b]^2$$

Suma kwadratów błędów $S(a, b)$ jest funkcją dwóch zmiennych. Z rachunku różniczkowego wiadomo, że aby wyznaczyć położenie ekstremum (w naszym przypadku minimum) funkcji dwóch zmiennych należy obliczyć jej pochodne cząstkowe i przyrównać je do zera. Ponieważ suma kwadratów błędów $S(a, b)$ jest funkcją złożoną

$$\frac{\partial S(a, b)}{\partial a} = 2 \sum_{i=1} [y_i - ax_i - b] \cdot -x_i = -2 \sum_{i=1} [y_i x_i - a{x_i}^2 - bx_i]$$

$$\frac{\partial S(a, b)}{\partial b} = 2 \sum_{i=1} [y_i - ax_i - b] \cdot -1 = -2 \sum_{i=1} [y_i - ax_i - b]$$

Warunek istnienia ekstremum (minimum)

$$\frac{\partial S(a,b)}{\partial a} = 0$$

$$\frac{\partial S(a,b)}{\partial b} = 0$$

Dodając (lub odejmując) powyższe równania stronami możemy również zapisać

$$\frac{\partial S(a,b)}{\partial a} \pm \frac{\partial S(a,b)}{\partial b} = 0$$

Aby wyznaczyć wartość współczynników a i b, konieczne jest rozwiązanie układu dwóch równań z dwiema niewiadomymi.

$$-2\sum_{i=1}[y_i x_i - a{x_i}^2 - bx_i] = 0$$

$$-2\sum_{i=1}[y_i - ax_i - b] = 0$$

W pierwszej kolejności pozbywamy się stałych współczynników (-2)

$$\sum_{i=1}[y_i x_i - a{x_i}^2 - bx_i] = 0$$

$$\sum_{i=1}[y_i - ax_i - b] = 0$$

$$\sum_{i=1} y_i x_i - \sum_{i=1} a{x_i}^2 - \sum_{i=1} bx_i = 0$$

$$\sum_{i=1} y_i - \sum_{i=1} ax_i - \sum_{i=1} b = 0$$

Oczywiście

$$\sum_{i=1} b = \sum_{i=1} bi^0 = b\sum_{i=1} i^0 = Nb$$

Wartości stałe wynosimy przed poszczególne sumy

$$\sum_{i=1} y_i x_i - a\sum_{i=1} {x_i}^2 - b\sum_{i=1} x_i = 0$$

$$\sum_{i=1} y_i - a\sum_{i=1} x_i - Nb = 0$$

Wprowadzamy oznaczenia pomocnicze

$$\sum_{i=1} y_i x_i = S_{y_i x_i}$$

$$\sum_{i=1} {x_i}^2 = S_{{x_i}^2}$$

$$\sum_{i=1} x_i = S_{x_i}$$

$$\sum_{i=1} y_i = S_{y_i}$$

Układ równań przyjmuje postać

$$S_{y_i x_i} - aS_{{x_i}^2} - bS_{x_i} = 0$$

$$S_{y_i} - aS_{x_i} - Nb = 0$$

$$S_{y_i x_i} - aS_{{x_i}^2} = bS_{x_i}$$

$$S_{y_i} - aS_{x_i} = Nb$$

Z drugiego równania wyznaczamy wartość wyrazu wolnego b.

$$b = \frac{S_{y_i} - aS_{x_i}}{N}$$

W dalszej kolejności zajmiemy się wyłącznie pierwszym równaniem.

$$S_{y_i x_i} - aS_{x_i{}^2} = \left(\frac{S_{y_i} - aS_{x_i}}{N}\right) S_{x_i}$$

$$S_{y_i x_i} - aS_{x_i{}^2} = \frac{S_{y_i} S_{x_i} - a\left(S_{x_i}\right)^2}{N}$$

$$NS_{y_i x_i} - aNS_{x_i{}^2} = \ S_{y_i} S_{x_i} - a\left(S_{x_i}\right)^2$$

$$-aNS_{x_i{}^2} + a\left(S_{x_i}\right)^2 = \ S_{y_i} S_{x_i} - \ NS_{y_i x_i}$$

$$-a\left[NS_{x_i{}^2} - \left(S_{x_i}\right)^2\right] = \ S_{y_i} S_{x_i} - \ NS_{y_i x_i}$$

$$a\left[NS_{x_i{}^2} - \left(S_{x_i}\right)^2\right] = \ NS_{y_i x_i} - \ S_{y_i} S_{x_i}$$

$$a = \frac{NS_{y_i x_i} - \ S_{y_i} S_{x_i}}{NS_{x_i{}^2} - \left(S_{x_i}\right)^2}$$

Teraz powracamy do sum w postaci jawnej

$$a = \frac{N\sum_{i=1} y_i x_i - \sum_{i=1} y_i \sum_{i=1} x_i}{N\sum_{i=1} x_i{}^2 - \left(\sum_{i=1} x_i\right)^2}$$

Do naszego równania wprowadzamy ukrytą jedynkę

$$\frac{N^2}{N^2} = 1$$

$$a = \frac{N\sum_{i=1} y_i x_i - \frac{N^2}{N^2}\sum_{i=1} y_i \sum_{i=1} x_i}{N\sum_{i=1} {x_i}^2 - \frac{N^2}{N^2}\left(\sum_{i=1} x_i\right)^2}$$

$$a = \frac{N\sum_{i=1} y_i x_i - N^2 \sum_{i=1} \frac{y_i}{N} \sum_{i=1} \frac{x_i}{N}}{N\sum_{i=1} {x_i}^2 - N^2\left(\sum_{i=1} \frac{x_i}{N}\right)^2}$$

Ale

$$\sum_{i=1} \frac{y_i}{N} = \langle y \rangle$$

$$\sum_{i=1} \frac{x_i}{N} = \langle x \rangle$$

A więc

$$a = \frac{N\sum_{i=1} y_i x_i - N^2 \langle x \rangle \langle y \rangle}{N\sum_{i=1}^{n} {x_i}^2 - N^2 \langle x \rangle^2} = \frac{N\left[\sum_{i=1} y_i x_i - N\langle x \rangle \langle y \rangle\right]}{N\left[\sum_{i=1} {x_i}^2 - N\langle x \rangle^2\right]}$$

$$a = \frac{\sum_{i=1} y_i x_i - N\langle x\rangle\langle y\rangle}{\sum_{i=1} {x_i}^2 - N\langle x\rangle^2}$$

Aby wyznaczyć wartość współczynnika kierunkowego a prostej regresji wystarczy obliczyć ile wynoszą następujące wielkości:

a) wartość średnia $\langle x\rangle$ ze zbioru zmiennych objaśniających x_i

b) wartość średnia $\langle y\rangle$ ze zbioru zmiennych objaśnianych y_i

c) sumę kwadratów zmiennych objaśniających

$$\sum_{i=1} {x_i}^2$$

d) sumę iloczynów odpowiadających sobie zmiennych objaśniających x_i i objaśnianych y_i

$$\sum_{i=1} y_i x_i = \sum_{i=1} x_i y_i$$

Wiemy, że:

$$b = \frac{S_{y_i} - aS_{x_i}}{N} = \frac{S_{y_i}}{N} - a\frac{S_{x_i}}{N} = \sum_{i=1} \frac{y_i}{N} - a\sum_{i=1} \frac{x_i}{N} = \langle y\rangle - a\langle x\rangle$$

$$b = -a\langle x\rangle + \langle y\rangle$$

Aby wyznaczyć wartość wyrazu wolnego b prostej regresji wystarczy obliczyć ile wynoszą następujące wielkości:

a) współczynnik kierunkowy a prostej regresji

b) wartość średnia $\langle x \rangle$ ze zbioru zmiennych objaśniających x_i

c) wartość średnia $\langle y \rangle$ ze zbioru zmiennych objaśnianych y_i

W metodzie najmniejszych kwadratów ważne znaczenie mają **obserwacje o dużej dźwigni**, które mają nietypowe wartości x_i i typowe wartości y_i (są znacząco oddalone od pozostałych punktów na osi poziomej) oraz **obserwacje odstające** (ang. *outliers*), które mają typowe wartości x_i i nietypowe wartości y_i (są znacząco oddalone od pozostałych punktów na osi pionowej). Zarówno **obserwacje o dużej dźwigni** jak i **obserwacje odstające** mogą mieć znaczny wpływ na **prostą regresji** (zmieniają wartości zarówno współczynnika kierunkowego a jak i wyrazu wolnego b). Jeśli posiadają taki wpływ, to nazywane są **obserwacjami wpływowymi** (ang. *influential*). To czy konkretna obserwacja jest wartością odstającą jest sprawą subiektywną - decyzję o identyfikacji obserwacji jako odstającej należy podjąć w oparciu o własne doświadczenie rachunkowe. Aby ocenić jakość dopasowania prostej regresji do zbioru analizowanych punktów (obserwacji empirycznych) stosowany jest **współczynnik determinacji** R^2 (iloraz wariancji wyznaczonej dla punktów, które leżą na prostej regresji $y(x_i)$ oraz wariancji wyznaczonej dla punktów doświadczalnych y_i), który przyjmuje wartość z przedziału $0 \leq R^2 \leq 1$. Oznaczenie współczynnik determinacji R^2 jest związane z jego definicją, w której każda z reszt (ang. *rest*) jest podniesiona do kwadratu. Im bliższa jedności jest wartość współczynnika determinacji R^2 tym lepsze jest dopasowanie prostej regresji do obserwacji empirycznych (doświadczalnych). W przypadku gdy $R^2 = 1$ wszystkie obserwacje empiryczne y_i leżą dokładnie na prostej regresji $y(x_i)$. Innymi słowy spełnione są równania $y_i = y(x_i)$, gdzie $i = 1,2,3, \dots, n$. Taka sytuacja jednak prawie nigdy się nie zdarza!

$$R^2 = \frac{\frac{1}{N}\sum_{i=1}[y(x_i) - \langle y \rangle]^2}{\frac{1}{N}\sum_{i=1}[y_i - \langle y \rangle]^2} = \frac{\sum_{i=1}[y(x_i) - \langle y \rangle]^2}{\sum_{i=1}[y_i - \langle y \rangle]^2}$$

$y(x_i)$ - wartość teoretyczna zmiennej objaśnianej, która leży na prostej regresji

$$y(x_i) = ax_i + b$$

y_i - empiryczna (doświadczalna) wartość zmiennej objaśnianej ($i = 1,2,3, \dots, n$)

$\langle y \rangle$ - średnia arytmetyczna obliczona na podstawie empirycznych wartości zmiennej objaśnianej

Powyżej mamy do czynienia z ilorazem dwóch wariancji (mianowniki ulegają skróceniu).

$$\frac{1}{N}\sum_{i=1}[y(x_i) - \langle y \rangle]^2 \quad \text{oraz} \quad \frac{1}{N}\sum_{i=1}[y_i - \langle y \rangle]^2$$

Zawsze prawdziwa jest nierówność

$$\sum_{i=1}[y(x_i) - \langle y \rangle]^2 \leq \sum_{i=1}[y_i - \langle y \rangle]^2$$

ponieważ odchylenia $y(x_i) - \langle y \rangle$ są znacznie mniejsze niż $y_i - \langle y \rangle$! Równość zachodzi jedynie w granicznym przypadku gdy prosta regresji przechodzi przez wszystkie punkty otrzymane na drodze doświadczalnej: $y(x_i) = y_i$

$$R^2 = \frac{\sum_{i=1}[y(x_i) - \langle y \rangle]^2}{\sum_{i=1}[y_i - \langle y \rangle]^2} = \frac{\sum_{i=1}[y_i - \langle y \rangle]^2}{\sum_{i=1}[y_i - \langle y \rangle]^2} = 1$$

Dopełnieniem do jedynki **współczynnika determinacji** R^2 jest **współczynnik zbieżności** φ^2 (gr. litera *phi*)

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

Współczynnik zbieżności φ^2 podobnie jak współczynnik determinacji R^2 przyjmuje wartości z przedziału $0 \leq \varphi^2 \leq 1$. Im bliższa zeru jest wartość współczynnika zbieżności φ^2 tym lepsze jest dopasowanie prostej regresji do obserwacji empirycznych.

Zwróćmy uwagę, że w metodzie najmniejszych kwadratów naszym celem nie jest znalezienie równania krzywej, która dokładnie przejdzie przez wszystkie punkty eksperymentalne a jedynie takiej, dla której rozbieżności pomiędzy zmierzonymi wartościami y_i a wartościami wyznaczonej funkcji liniowej $y(x_i)$ będą najmniejsze. W metodzie najmniejszych kwadratów nie jest wymagany stały odstęp pomiędzy poszczególnymi wartościami x_i zmiennej objaśniającej.

Jakie są ograniczenia w stosowaniu metody najmniejszych kwadratów?

a) pomiędzy zmiennymi losowymi X i Y powinna występować zależność, która jest liniowa.

b) rozkład reszt $r = y_i - y(x_i)$, a więc różnic pomiędzy wartościami eksperymentalnymi y_i a wartościami teoretycznymi $y(x_i)$, które leżą na prostej regresji powinien być zbliżony do rozkładu normalnego lub być rozkładem normalnym.

c) rozproszenie punktów dookoła prostej regresji powinno być w miarę równomierne.