GHOST
Group of Horribly Optimistic Statisticians

Intro to ML

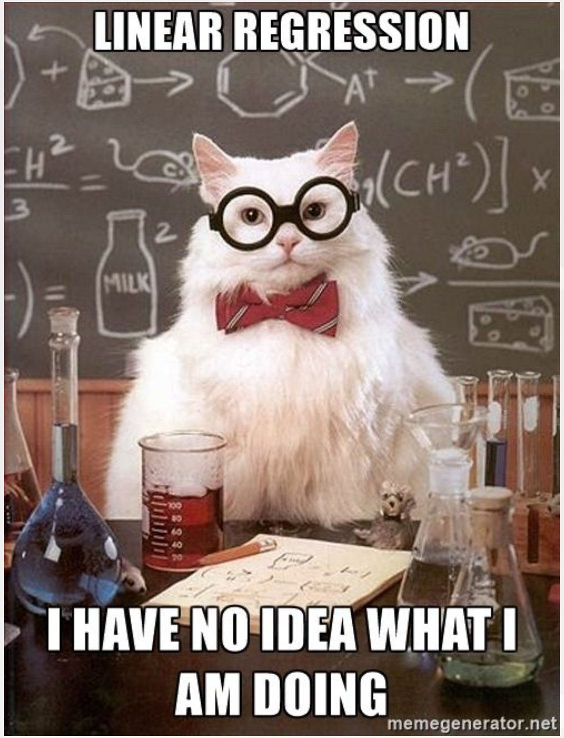
#3 Regresja

Martyna Stasiak | 05.05.2025

GHOST
Group of Horribly Optimistic Statisticians

Agenda

1. Czym jest regresja
2. Zastosowania
3. Regresja Liniowa
4. Regresja Wielomianowa
5. Regresja Logistyczna

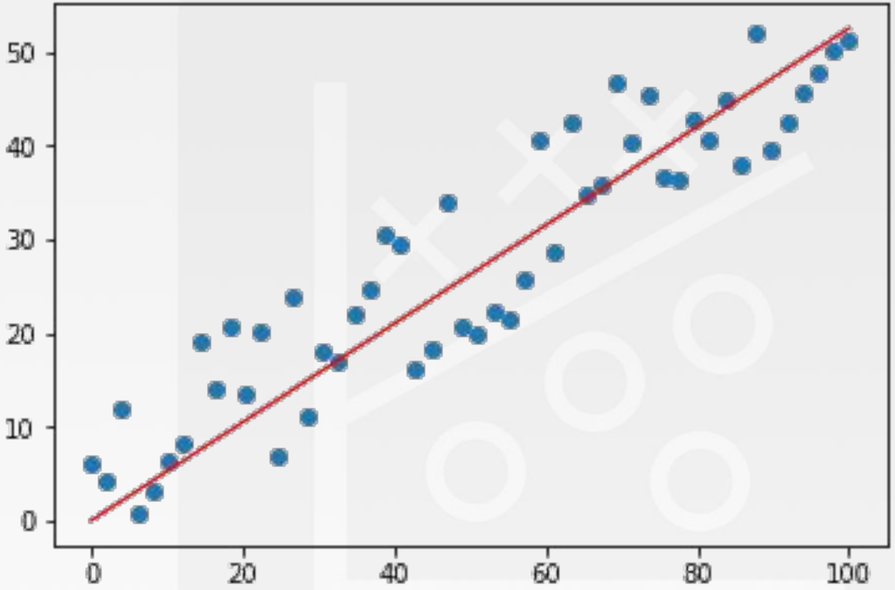


GHOST

Group of Horribly Optimistic Statisticians

Czym jest regresja?

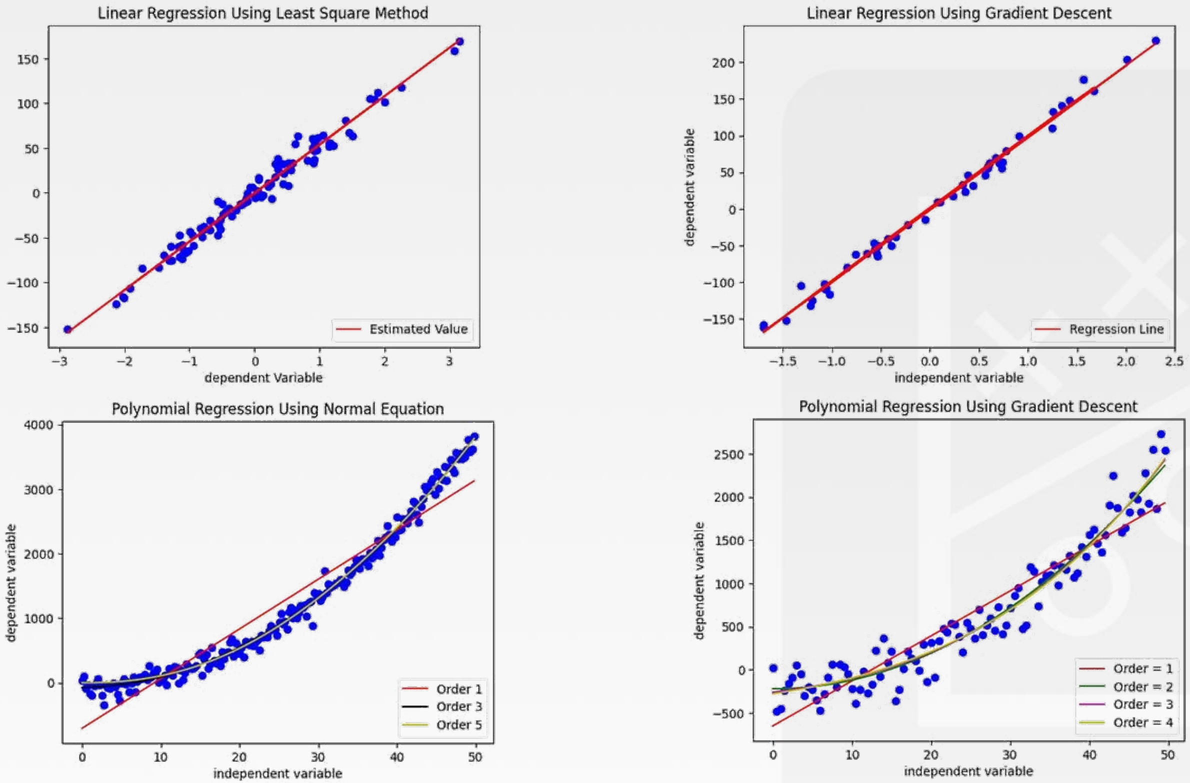
Regresja pozwala nam przewidzieć ciągłe wartości (continuous values) na podstawie dostarczonych danych. Wykorzystuje ona rzeczywiste wartości do przewidywania danych ilościowych, takich jak dochód, wzrost, waga, wyniki czy prawdopodobieństwo.



GHOST

Group of Horribly Optimistic Statisticians

Regression Analysis



GHOST
Group of Horribly Optimistic Statisticians

Zastosowania Regresji

1. Prognozowanie:

Regresja pozwala przewidywać wartości ciągłe na podstawie wzorców w danych.

Przykłady zastosowań:

- *Prognoza cen akcji na rynku giełdowym w oparciu o dane historyczne.*
- *Szacowanie sprzedaży produktów w oparciu o dane z poprzednich okresów.*
- *Przewidywanie popytu na energię elektryczną w różnych porach dnia i roku.*

GHOST

Group of Horribly Optimistic Statisticians

Zastosowania Regresji

2. Analiza trendów:

Umożliwia zrozumienie, jak zmienne są ze sobą powiązane i w jaki sposób jedna zmienna wpływa na drugą.

Przykłady:

- *Analiza zależności między ceną produktu a jego sprzedażą (np. wyższa cena -> mniejszy popyt).*
- *Badanie, jak wzrost temperatury wpływa na zużycie energii klimatyzacyjnej.*

Zastosowania Regresji

3. Modelowanie zjawisk:

Regresja jest używana do budowy modeli wyjaśniających zjawiska w świecie rzeczywistym.

Przykłady:

- *Analiza skutków zmian polityki podatkowej na dochody budżetowe.*
- *Symulacja wpływu różnego rodzaju reklam na decyzje zakupowe konsumentów.*

Zastosowania Regresji

4. Redukcja złożoności:

Pomaga uprościć skomplikowane zależności w danych, tworząc modele wyjaśniające kluczowe wzorce.

Przykłady:

- *Budowa prostego modelu przewidującego masę ciała na podstawie wzrostu i wieku.*
- *Tworzenie modeli wyjaśniających zachowanie konsumentów na podstawie kilku kluczowych zmiennych.*

GHOST
Group of Horribly Optimistic Statisticians

Rodzaje regresji

1. Regresja Liniowa: Modelowanie liniowej zależności między zmiennymi.
2. Regresja Wielomianowa: Modelowanie nieliniowych zależności poprzez wielomiany.
3. Regresja Logistyczna: Modelowanie prawdopodobieństwa wystąpienia zdarzenia binarnego.
4. Regresja Ridge i Lasso: Techniki regularizacji zapobiegające przeuczeniu modelu.

Regresja Liniowa

Regresja liniowa przewiduje zależność między dwiema zmiennymi, zakładając, że mają one liniową relację (prostoliniową). Jej celem jest znalezienie najlepszej linii, która minimalizuje różnice między wartościami rzeczywistymi a przewidywanymi.

Regresja Liniowa

Regresja liniowa przewiduje zależność między dwiema zmiennymi, zakładając, że mają one liniową relację (prostoliniową).
Jej celem jest znalezienie najlepszej linii, która minimalizuje różnice między wartościami rzeczywistymi a przewidywanymi.

Regresja liniowa znajduje zastosowanie w takich dziedzinach jak ekonomia czy finanse, pomagając w analizie i prognozowaniu trendów danych. Może obejmować jedną zmienną niezależną (simple linear regression) lub wiele zmiennych (multiple linear regression).

GHOST
Group of Horribly Optimistic Statisticians

Regresja Liniowa

Simple Linear Regression

W prostej regresji liniowej występuje jedna zmienna niezależna (predyktor) i jedna zmienna zależna (output).

Model szacuje nachylenie i punkt przecięcia linii najlepszego dopasowania, która reprezentuje relację między tymi zmiennymi. Nachylenie wskazuje, jak zmienia się zmienna zależna wraz ze zmianą zmiennej niezależnej o jednostkę, podczas gdy punkt przecięcia reprezentuje przewidywaną wartość zmiennej zależnej, gdy zmienna niezależna wynosi zero.

GHOST

Group of Horribly Optimistic Statisticians

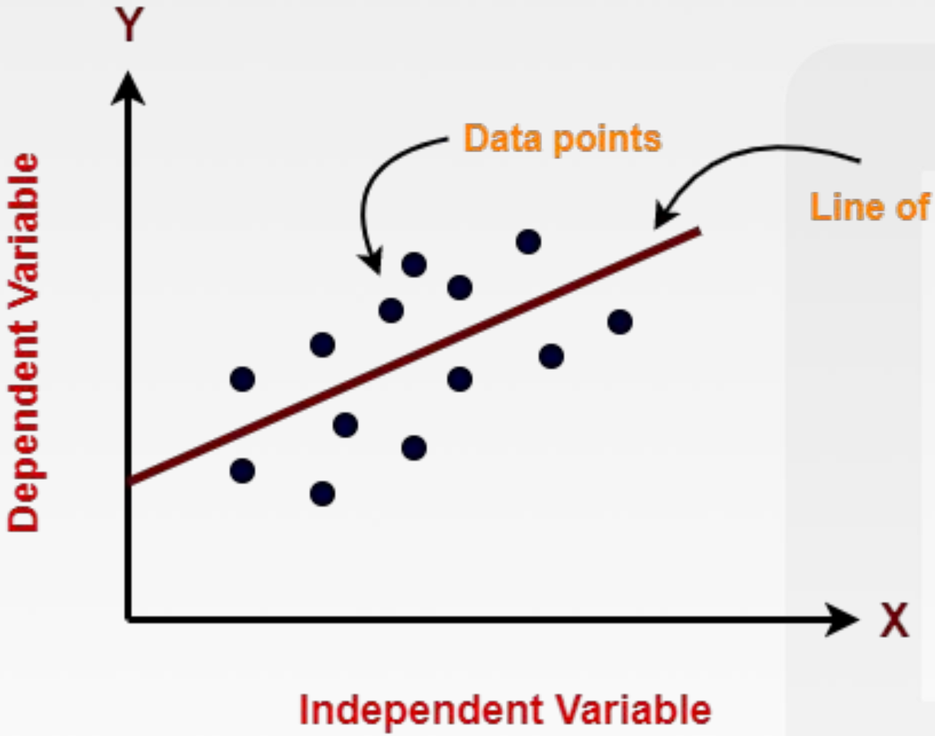
Regresja Liniowa

GHOST
Group of Horribly Optimistic Statisticians

Czym jest linia najlepszego dopasowania?

W prostych słowach, linia najlepszego dopasowania to linia, która najlepiej dopasowuje się do punktów na wykresie rozrzutu (scatter plot). Matematycznie linię tę wyznacza się poprzez minimalizację sumy kwadratów reszt (ang. Residual Sum of Squares, RSS).

Minimalizacja RSS oznacza, że różnice między rzeczywistymi wartościami a wartościami przewidywanymi przez model są jak najmniejsze.


Y
Data points
Line of Regression
Dependent Variable
X
Independent Variable

GHOST
Group of Horribly Optimistic Statisticians

Simple Linear Regression

Aby obliczyć linię najlepszego dopasowania, regresja liniowa korzysta z tradycyjnej postaci równania prostoliniowego, która wygląda następująco:

$$Yi=\beta0+\beta1Xi$$

- Yi: zmienna zależna (output),
- β0: stała (punkt przecięcia z osią Y),
- β1: nachylenie linii (slope),
- Xi: zmienna niezależna (predyktor).

Algorytm więc wyjaśnia liniową zależność między zmianną zależną- naszym outputem y, a zmienną niezależną - predyktorem x za pomocą powyższego równania linii prostej. Linia ta reprezentuje najlepsze dopasowanie do danych, minimalizując różnice między wartościami rzeczywistymi a przewidywanymi przez model.

GHOST
Group of Horribly Optimistic Statisticians

Simple Linear Regression

Przykład:

Jeśli chcesz przewidzieć cenę domu (Y) wyłącznie na podstawie powierzchni domu (X1), równanie regresji liniowej przyjmuje postać:

$$\text{Cena domu} = \beta_0 + \beta_1 \cdot \text{Powierzchnia} + \varepsilon$$

GHOST
Group of Horribly Optimistic Statisticians

Jak regresja liniowa znajduje linię najlepszego dopasowania?

Celem algorytmu regresji liniowej jest znalezienie najlepszych wartości dla $\beta 0$ (punkt przecięcia) i $\beta 1$ (nachylenie), aby wyznaczyć linię najlepszego dopasowania. Linia najlepszego dopasowania to taka, która minimalizuje błąd, co oznacza, że różnica między wartościami przewidywanymi a rzeczywistymi powinna być jak najmniejsza.

Błąd losowy - Reszty

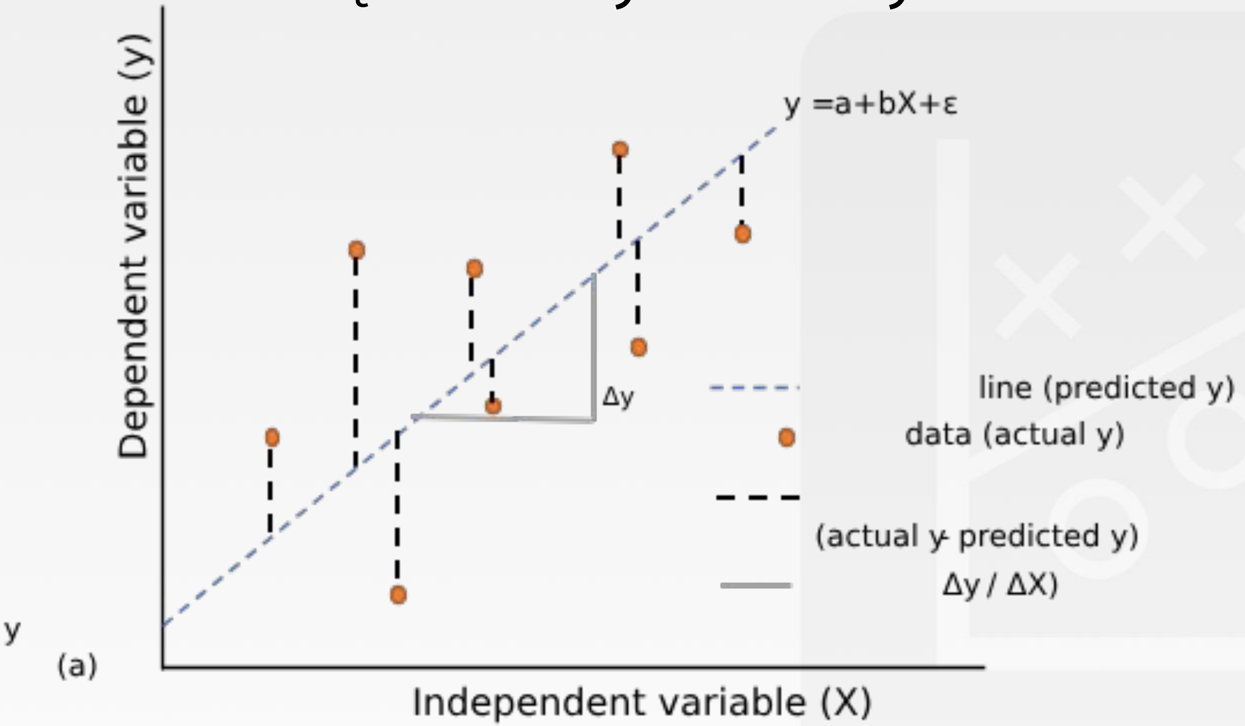
W regresji różnica między zaobserwowaną wartością zmiennej zależnej (y_i) a wartością przewidywaną przez model ($y_{predicted}$) nazywana jest resztą.

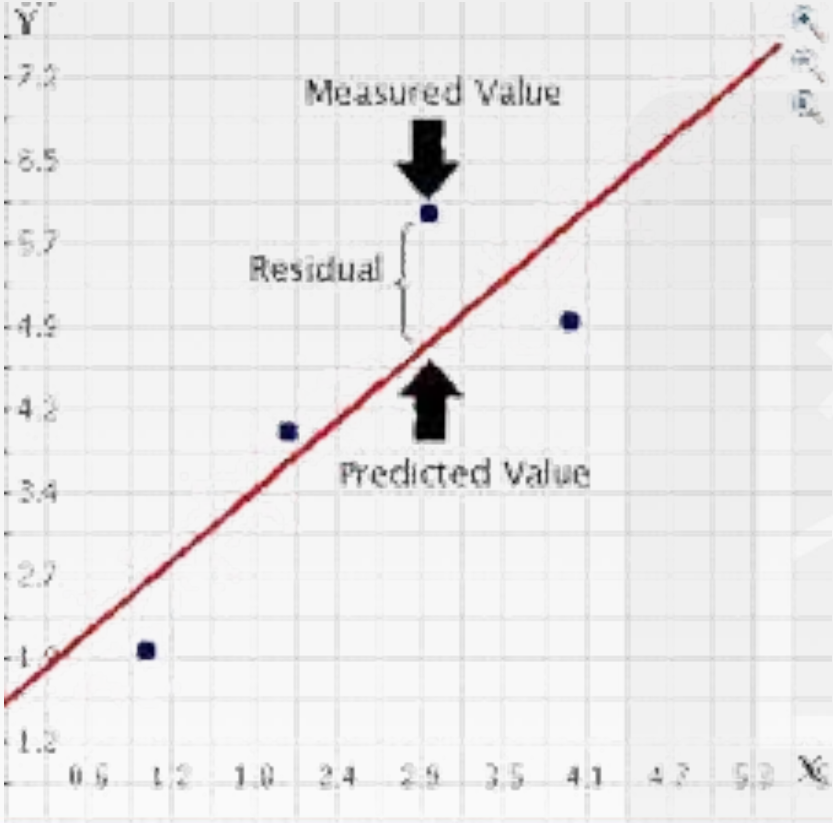
$$\varepsilon_i = y_{predicted} - y_i$$

- $y_{predicted}=\beta_0+\beta_1 X_i$: wartość przewidywana przez model,
- y_i: rzeczywista wartość zmiennej zależnej.

Reszty reprezentują błąd w przewidywaniu wartości przez model i są kluczowym elementem oceny jakości dopasowania modelu do danych.

Błąd losowy - Reszty




Y
Measured Value
Residual
Predicted Value
X

Funkcja kosztu w regresji liniowej

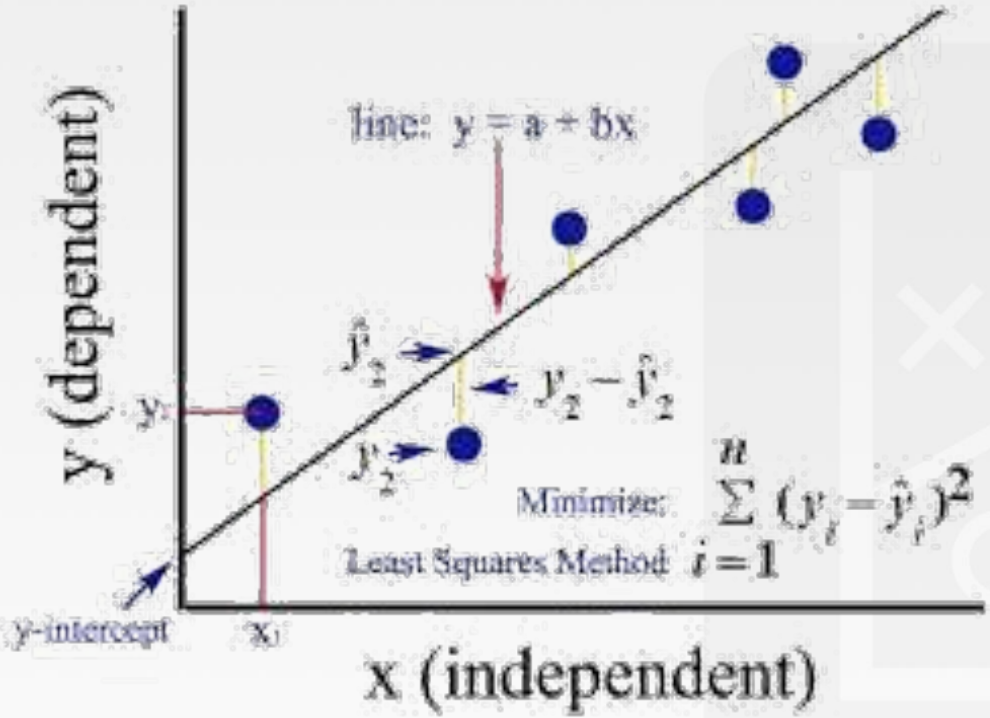
Funkcja kosztu służy do wyznaczania optymalnych wartości β0 i β1, które zapewniają linię najlepszego dopasowania do danych.

W regresji liniowej najczęściej stosuje się funkcję kosztu opartą na średnim błędzie kwadratowym (Mean Squared Error, MSE). MSE to średnia wartość kwadratów różnic między wartością przewidywaną (ypredicted) a rzeczywistą (yi).

Obliczenie MSE wykorzystuje równanie prostej regresji y=mx+b (odpowiednik y=β0+β1x) i jest dane wzorem:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (\mathrm{B}1x_i + \mathrm{B}0))^2$$

Funkcja kosztu w regresji liniowej



Metryki oceny regresji liniowej

Jakość każdego modelu regresji liniowej można ocenić za pomocą różnych metryk, które mierzą, jak dobrze model przewiduje rzeczywiste wartości. Metryki te dostarczają informacji o tym, na ile dokładnie model odzwierciedla dane obserwowane.

Najczęściej stosowane metryki to:

1. Współczynnik determinacji (R-squared, R2)
2. Pierwiastek ze średniego błędu kwadratowego (Root Mean Squared Error, RMSE)
3. Standardowy błąd reszt (Residual Standard Error, RSE)

GHOST

Group of Horribly Optimistic Statisticians

Współczynnik determinacji – (R-squared, R2)

Współczynnik determinacji (R2) to miara, która określa, jaka część zmienności zmiennej zależnej (y) jest wyjaśniona przez model. Wartość R2 zawsze mieści się w przedziale od 0 do 1, gdzie:

1. R2=1 oznacza idealne dopasowanie modelu do danych,
2. R2=0 wskazuje, że model nie wyjaśnia żadnej zmienności zmiennej zależnej.

Im wyższe R2, tym lepiej model pasuje do danych.

GHOST

Group of Horribly Optimistic Statisticians

Współczynnik determinacji – (R-squared, R2)

Wzór matematyczny: R2=1–RSS/TSS

Gdzie:

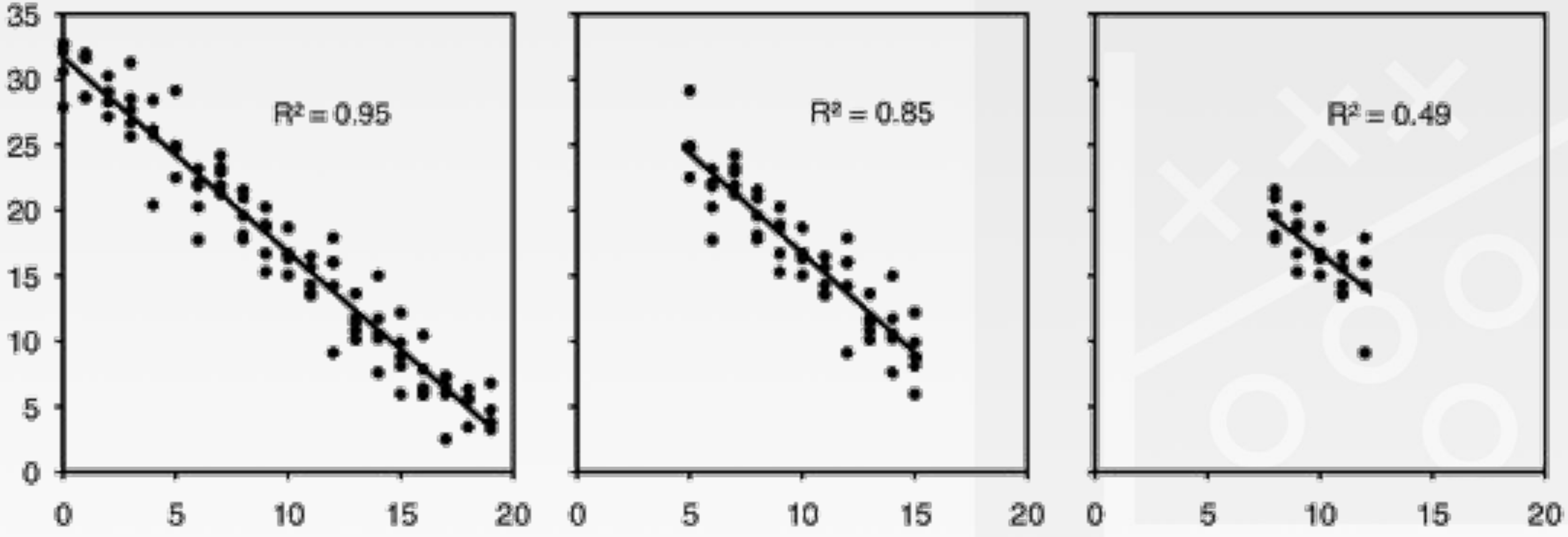
- RSS (Residual Sum of Squares): suma kwadratów reszt, czyli miara różnicy między przewidywanymi wartościami (ypredicted) a rzeczywistymi (yi). Wyraża się jako: **RSS=∑(yi–ypredicted)**2**
- TSS (Total Sum of Squares): suma kwadratów odchyleń rzeczywistych wartości (yi) od średniej wartości (y¯). Wyraża się jako: **TSS=∑(yi–y¯)**2**
- RSS: mierzy różnicę między wartościami przewidywanymi a rzeczywistymi. Im mniejszy RSS, tym lepsze dopasowanie modelu.
- TSS: mierzy całkowitą zmienność danych w odniesieniu do ich średniej.

GHOST
Group of Horribly Optimistic Statisticians

Współczynnik determinacji – (R-squared, R2)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Współczynnik determinacji – (R-squared, R2)



Pierwiastek ze średniego błędu kwadratowego (Root Mean Square Error, RMSE)

Pierwiastek ze średniego błędu kwadratowego (RMSE) to miara, która określa, jak dobrze model dopasowuje się do danych. Jest to pierwiastek z wariancji reszt, czyli różnic między wartościami rzeczywistymi a przewidywanymi. RMSE mierzy bezwzględne dopasowanie modelu do danych, wskazując, jak blisko wartości obserwowane są względem wartości przewidywanych.

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n}\left(y_i^{Actual} - y_i^{Predicted}\right)^2 / n}$$

GHOST

Group of Horribly Optimistic Statisticians

Standardowy błąd reszt (Residual Standard Error, RSE)

Aby oszacowanie błędu reszt było unbiased, należy podzielić sumę kwadratów reszt (Residual Sum of Squares, RSS) przez liczbę stopni swobody, a nie przez całkowitą liczbę punktów danych w modelu. Ten wynik nazywamy standardowym błędem reszt (RSE).

$$RSE = \sqrt{\frac{RSS}{df}} = \sqrt{\sum_{i=1}^{n}\left(y_i^{Actual} - y_i^{Predicted}\right)^2 \Big/ (n-2)}$$

GHOST
Group of Horribly Optimistic Statisticians

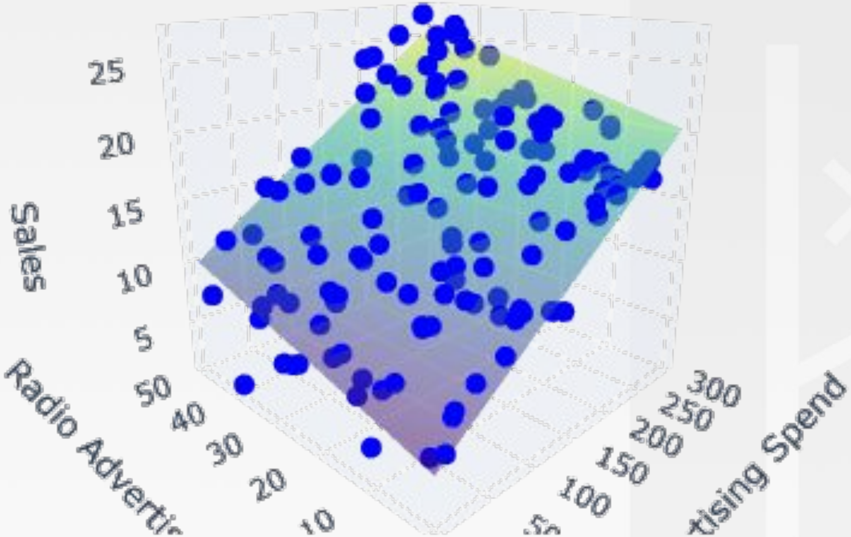
Multiple Linear Regression

Regresja wielokrotna to technika statystyczna, która pozwala zrozumieć relację między jedną zmienną zależną a wieloma zmiennymi niezależnymi. Jest rozszerzeniem prostej regresji liniowej, które pozwala uwzględniać wpływ więcej niż jednej zmiennej wejściowej.

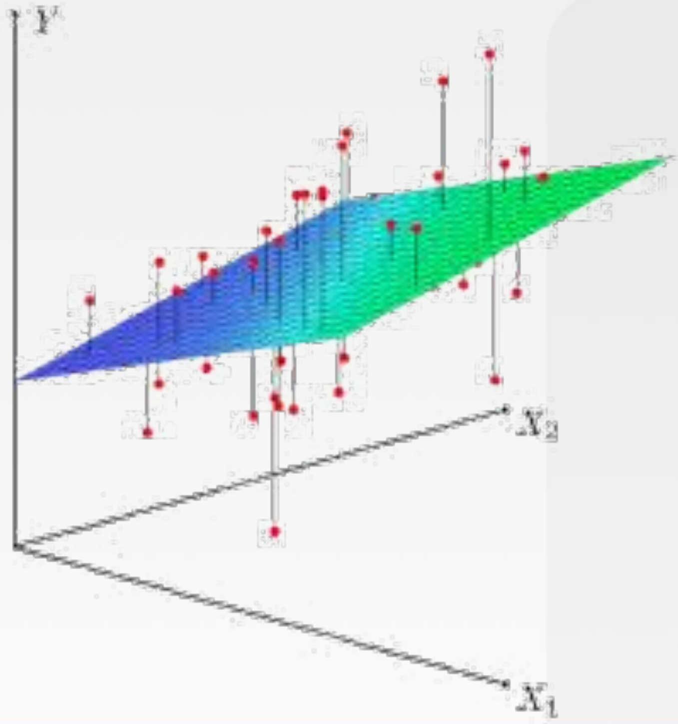
GHOST

Group of Horribly Optimistic Statisticians

Multiple Linear Regression



Multiple Linear Regression



Multiple Linear Regression

Formuła regresji wielokrotnej jest podobna do równania prostej regresji liniowej, z tą różnicą, że zamiast jednego współczynnika ($\beta 1$) dla zmiennej niezależnej, mamy oddzielne współczynniki ($\beta 1, \beta 2, \ldots, \beta p$) dla każdej zmiennej. Równanie wygląda następująco:

$$Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta p Xp + \varepsilon$$

Gdzie:

- Y: zmienna zależna (wynikowa),
- $X1, X2, \ldots, Xp$: zmienne niezależne (predyktory),
- $\beta 0$: wyraz wolny (przecięcie z osią Y),
- $\beta 1, \beta 2, \ldots, \beta p$: współczynniki regresji dla każdej zmiennej $X1, X2, \ldots, Xp$, które określają wpływ tych zmiennych na Y,
- ε: reszta (błąd przewidywań).

Multiple Linear Regression

Przykład:

Jeśli przewidujesz cenę domu (Y) na podstawie takich zmiennych jak 1) powierzchnia domu (X1), 2) liczba pokoi (X2) i 3) lokalizacja (X3), równanie może wyglądać tak:

$$\text{Cena domu} = \beta 0 + \beta 1 \cdot \text{Powierzchnia} + \beta 2 \cdot \text{Liczba pokoi} + \beta 3 \cdot \text{Lokalizacja} + \varepsilon$$

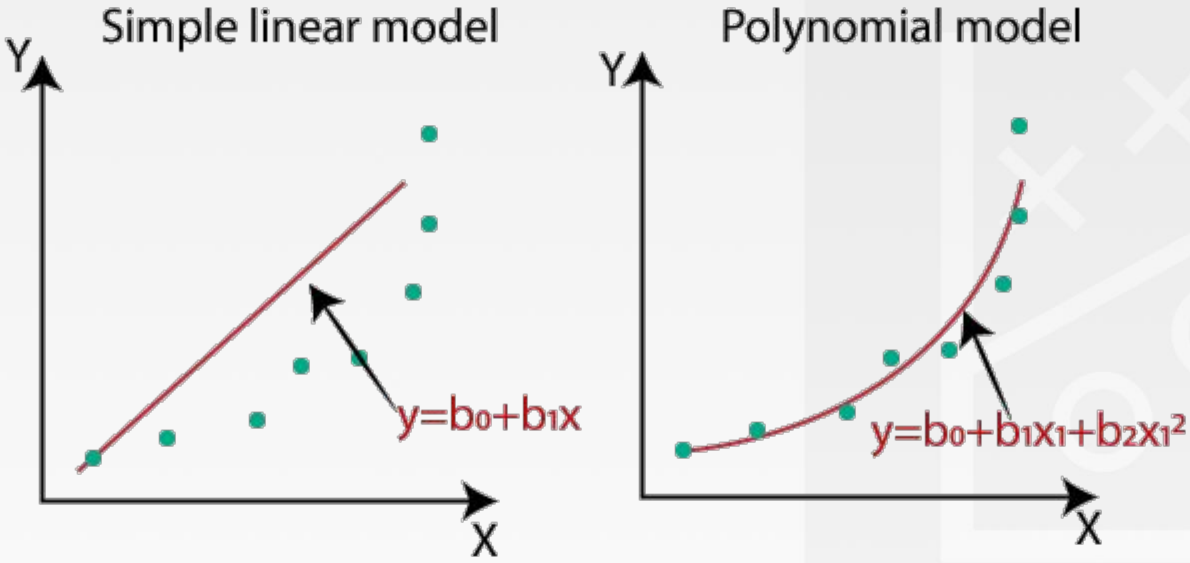
Regresja wielomianowa (Polynomial Regression)

Regresja wielomianowa to rozszerzenie regresji liniowej, które pozwala modelować nieliniowe zależności między zmienną zależną (y) a zmiennymi niezależnymi (x).

W przeciwieństwie do regresji liniowej, gdzie zależność opisana jest linią prostą, regresja wielomianowa używa krzywych, które mogą lepiej dopasować się do bardziej złożonych wzorców w danych.

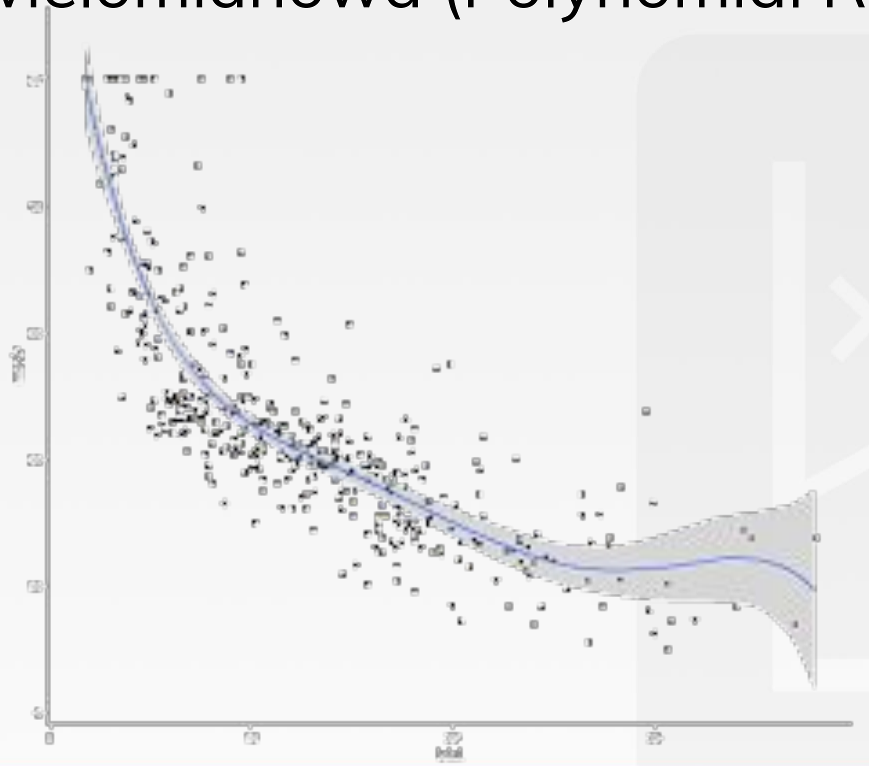
GHOST
Group of Horribly Optimistic Statisticians

Regresja wielomianowa (Polynomial Regression)



GHOST
Group of Horribly Optimistic Statisticians

Regresja wielomianowa (Polynomial Regression)



Regresja wielomianowa

Przykład:

Załóżmy, że chcemy przewidzieć cenę domu (y) na podstawie jego powierzchni (x), gdzie zależność nie jest liniowa. Równanie regresji wielomianowej może wyglądać tak:

*Cena domu=β0+β1x+β2x**2+β3x**3+ε*

Na wykresie dane punktowe są dopasowane za pomocą krzywej, która lepiej opisuje wzorce niż linia prosta.

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

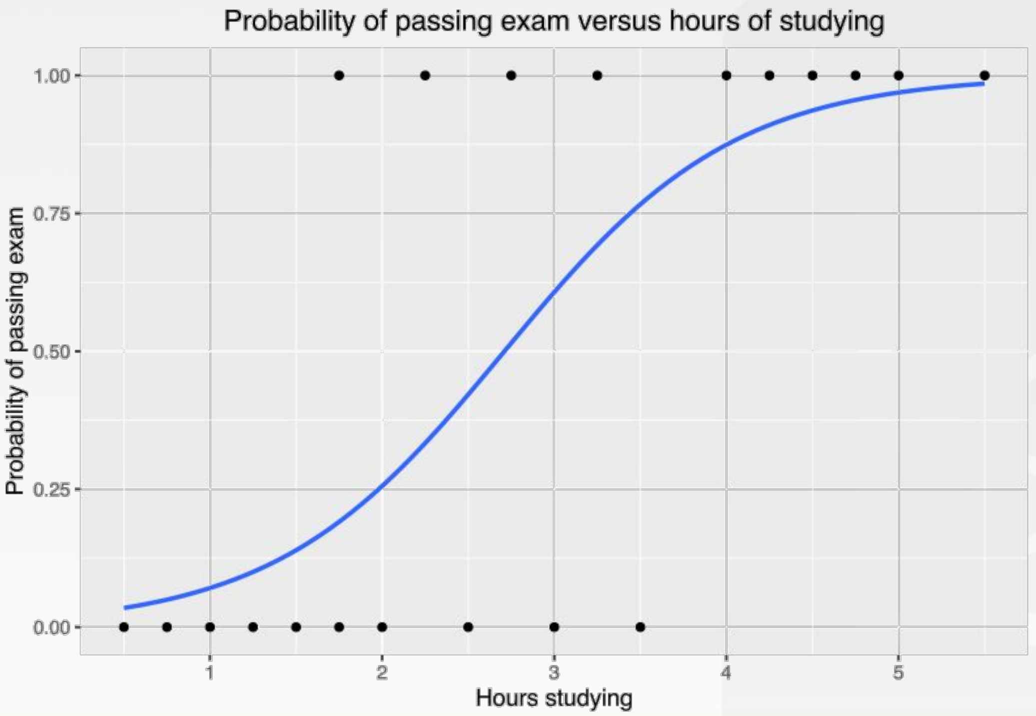
Regresja logistyczna

Regresja logistyczna to technika statystyczna i metoda uczenia maszynowego używana do przewidywania prawdopodobieństwa wystąpienia zdarzenia binarnego (np. tak/nie, 0/1)

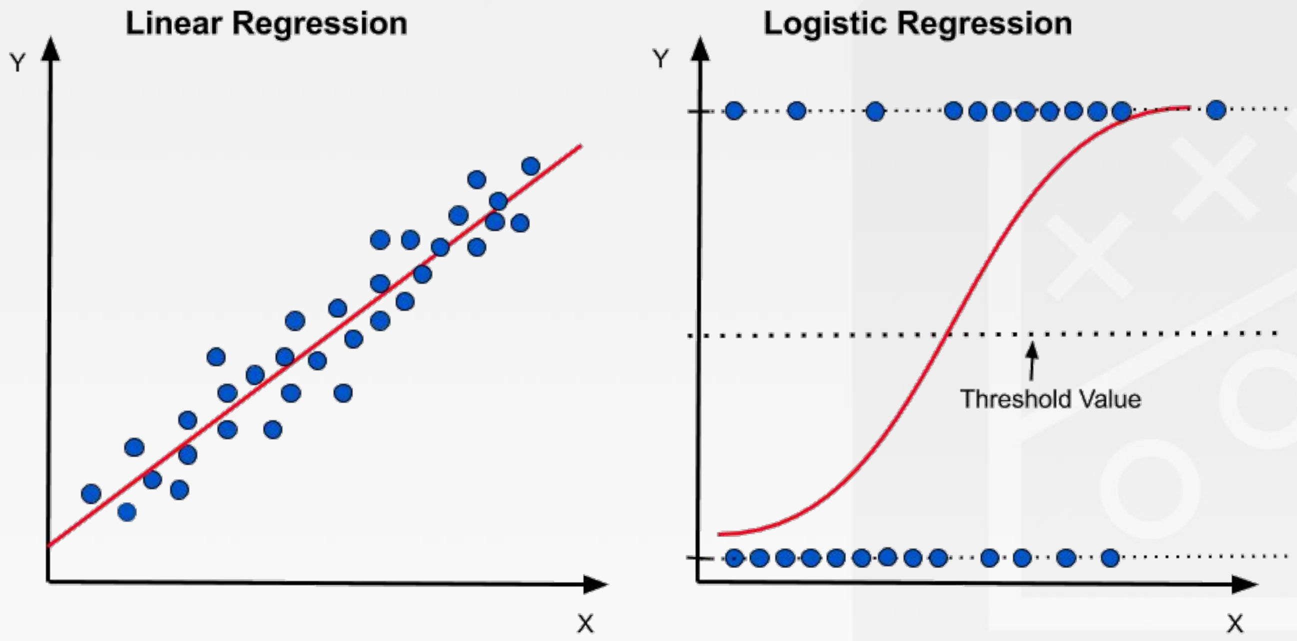
W przeciwieństwie do regresji liniowej, która przewiduje wartości ciągłe, regresja logistyczna przewiduje wynik w formie prawdopodobieństwa (wartości od 0 do 1).

Wynik modelu można interpretować jako prawdopodobieństwo wystąpienia zdarzenia.

Regresja logistyczna



Regresja logistyczna



Regresja logistyczna

Przykład:

Przewidywanie szans na zdanie egzaminu (y) na podstawie liczby godzin nauki (X):
Równanie modelu logistycznego:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- *P(y=1 |X): Prawdopodobieństwo zdania egzaminu (y=1).*
- *Jeśli student uczył się 5 godzin (X=5) i model przewiduje P(y=1 |X)=0.85, to prawdopodobieństwo zdania egzaminu wynosi 85%.*

Regresja logistyczna

Wykorzystanie:

- *Diagnoza medyczna:*
  - *Przewidywanie, czy pacjent ma określoną chorobę (np. y=1y=1: chory, y=0y=0: zdrowy).*
- *Analiza churnu klientów:*
  - *Przewidywanie, czy klient zrezygnuje z usługi.*
- *Klasyfikacja spamu:*
  - *Klasyfikacja wiadomości e-mail jako spam (y=1y=1) lub nie-spam (y=0y=0).*
- *Modelowanie ryzyka kredytowego:*
  - *Przewidywanie, czy klient spłaci kredyt.*

Regresja logistyczna

ZALETY	WADY
Prostota Łatwa do zrozumienia i wdrożenia, szczególnie dla problemów binarnej klasyfikacji.	**Założenie liniowości** Zakłada liniową relację między zmiennymi niezależnymi a logitową wartością wyniku, co może być ograniczeniem przy złożonych zależnościach.
Interpretowalność Współczynniki (β) wskazują, jaki wpływ mają zmienne wejściowe na prawdopodobieństwo wyniku.	**Wrażliwość na wartości odstające** Punkty odstające mogą silnie wpłynąć na wyniki.
Efektywność Działa dobrze przy dużych zbiorach danych z liniowymi zależnościami.	**Brak obsługi wieloklasowej** Standardowa regresja logistyczna działa dla problemów binarnych.

GHOST

Group of Horribly Optimistic Statisticians

Dziękuję za uwagę!

GHOST

Group of Horribly Optimistic Statisticians