Michał Leśniak 195642

Poniedziałek 17:45
20.12.2021r.

Komputerowa analiza danych

Zadanie 2

Cel

Dla podanych zestawów danych i proponowanych modeli oszacuj wartości parametrów a..f, które minimalizują średni błąd kwadratowy modelowanej funkcji w podanych punktach.

Wyniki

Poniższe wyniki zostały uzyskane przy użyciu programu napisanego w języku Python przez autora tego sprawozdania z wykorzystaniem bibliotek matplotlib, numpy oraz udostępnionego skryptu chi2_normality.py.

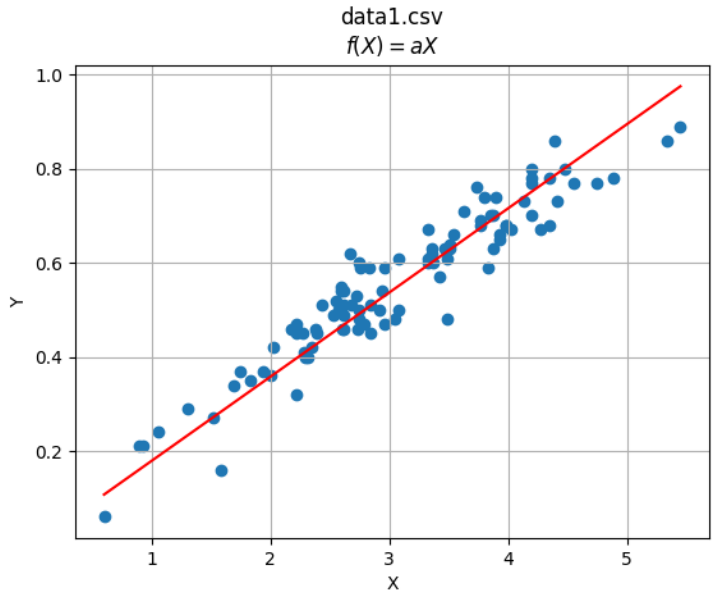
Model $f(X)=aX$

W celu przygotowania danych do zastosowania prostej regresji liniowej, wszystkie wartości z kolumny X oraz kolumny z wartościami w zestawach danych zostały wczytane do odpowiednio listy X oraz listy Y. Parametr a został obliczony ze wzoru $a=\frac{\overline{XY}}{\overline{X^2}}$

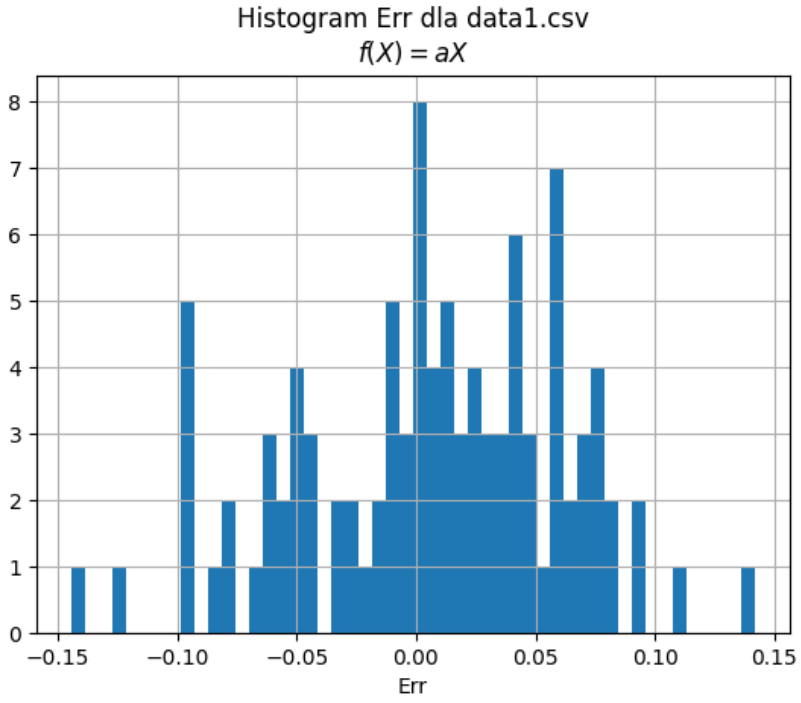
Tabela 1: Wyznaczone wartości parametrów dla modelu f(X) = aX i zestawu danych data1.csv

Parametr	**Wartość**
a	0,1811369929956019
średni błąd kwadratowy	0,003088987799646239
największa wartość odchylenia	0,1445985790725529
współczynnik R^2	0,8827351122800478

Wykres 1: Wykres przedstawiający modelowaną funkcję f(X) = aX na tle punktów z zestawu danych data1.csv



Histogram 1: Histogram odchyleń wartości funkcji f(X) = aX od danych z zestawu danych data1.csv



Test hipotezy statystycznej dla $f(X)=aX$ i zestawu danych data1.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

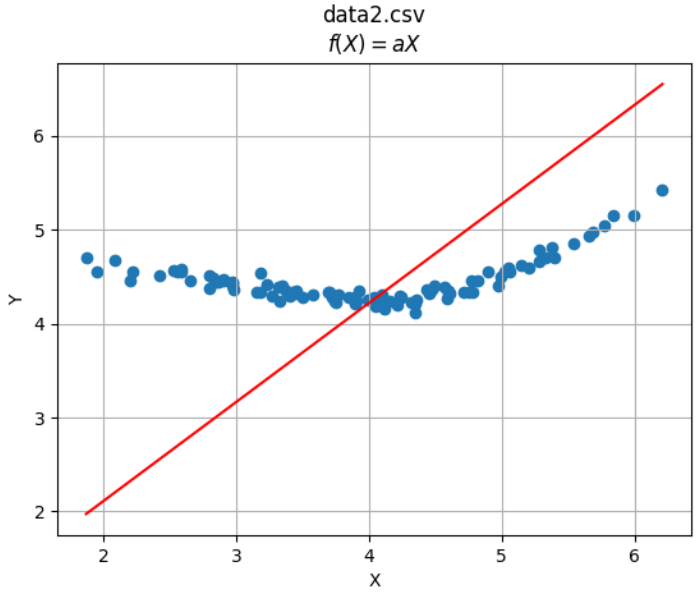
Otrzymana p-wartość: 0,0586755

Nie udało się odrzucić hipotezy.

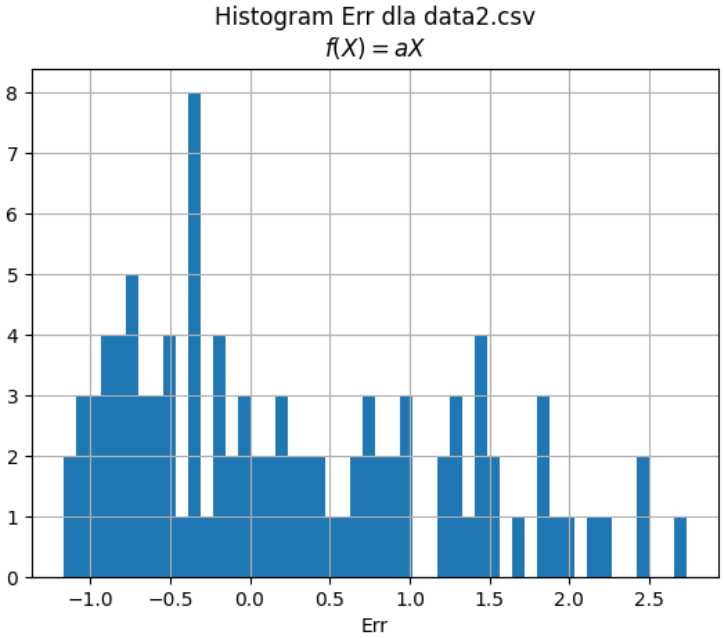
Tabela 2: Wyznaczone wartości parametrów dla modelu f(X) = aX i zestawu danych data2.csv

Parametr	Wartość
a	1,0550468408004456
średni błąd kwadratowy	1,0435240083167863
największa wartość odchylenia	2,7370624077031662
współczynnik R^2	-17,859854134254988

Wykres 2: Wykres przedstawiający modelowaną funkcję f(X) = aX na tle punktów z zestawu danych data2.csv



Histogram 2: Histogram odchyleń wartości funkcji f(X) = aX od danych z zestawu danych data2.csv



Test hipotezy statystycznej dla $f(X)=aX$ i zestawu danych data2.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

Otrzymana p-wartość: 0,0038674

Hipoteza odrzucona. Nie wydaje się by błędy miały rozkład normalny.

Ocena przydatności

Biorąc pod uwagę współczynniki R^2 oraz testy zgodności χ^2 Pearsona, dopasowanie modelu $f(X)=aX$ do zestawu danych data1.csv można uznać za sukces (R^2 powyżej 85% oraz p-wartość większa od ustalonego poziomu istotności). Wizualnie również dopasowanie wygląda poprawnie. Inaczej jest w przypadku zestawu danych data2.csv, gdzie współczynnik R^2 posiada wartość poniżej 0, a test zgodności χ^2 Pearsona zaprzeczył by błędy miały rozkład normalny. Wizualnie również regresja odbiega od krotek.

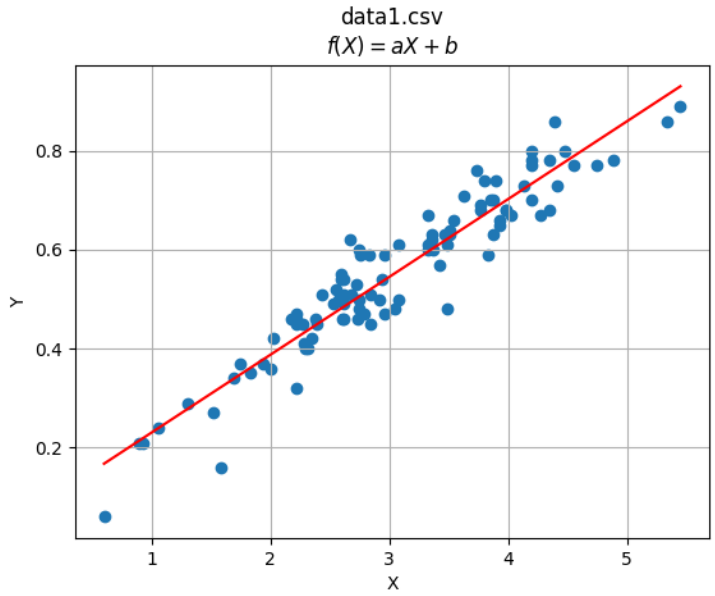
Model $f(X)=aX+b$

W celu przygotowania danych do zastosowania prostej regresji liniowej, wszystkie wartości z kolumny X oraz kolumny z wartościami w zestawach danych zostały wczytane do odpowiednio listy X oraz listy Y. Parametr a został obliczony ze wzoru $a=\frac{Cov(X,Y)}{VarX}$, natomiast parametr b: $b=\overline{Y}-a\,\overline{X}$.

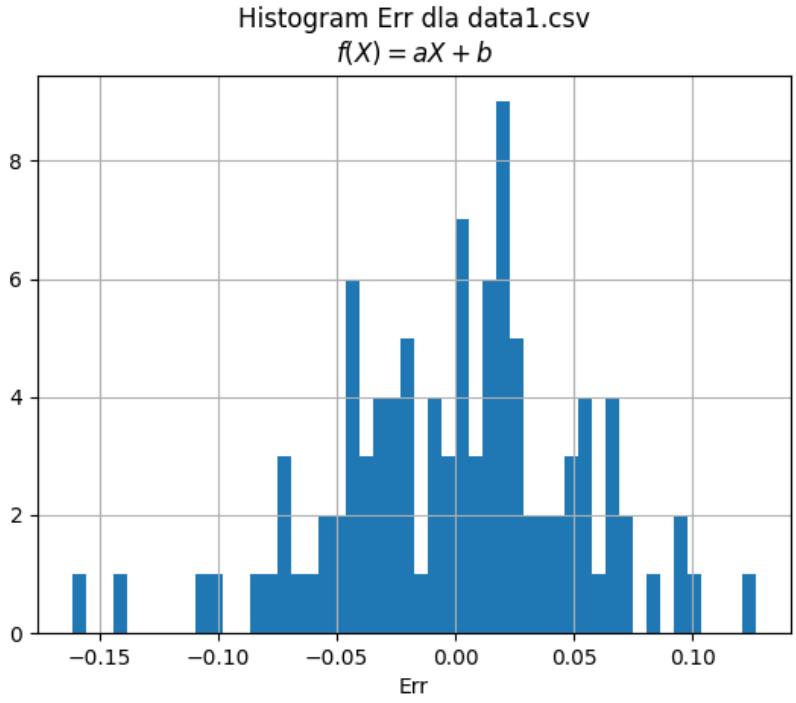
Tabela 3: Wyznaczone wartości parametrów dla modelu f(X) = aX + b i zestawu danych data1.csv

Parametr	**Wartość**
a	0,15731913673058445
b	0,07310890980547108
średni błąd kwadratowy	0,0026022655146449476
największa wartość odchylenia	0,16167314583979456
współczynnik R^2	0,8997740904850968

Wykres 3: Wykres przedstawiający modelowaną funkcję f(X) = aX + b na tle punktów z zestawu danych data1.csv



Histogram 3: Histogram odchyleń wartości funkcji f(X) = aX + b od danych z zestawu danych data1.csv



Test hipotezy statystycznej dla $f(X)=aX+b$ i zestawu danych data1.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

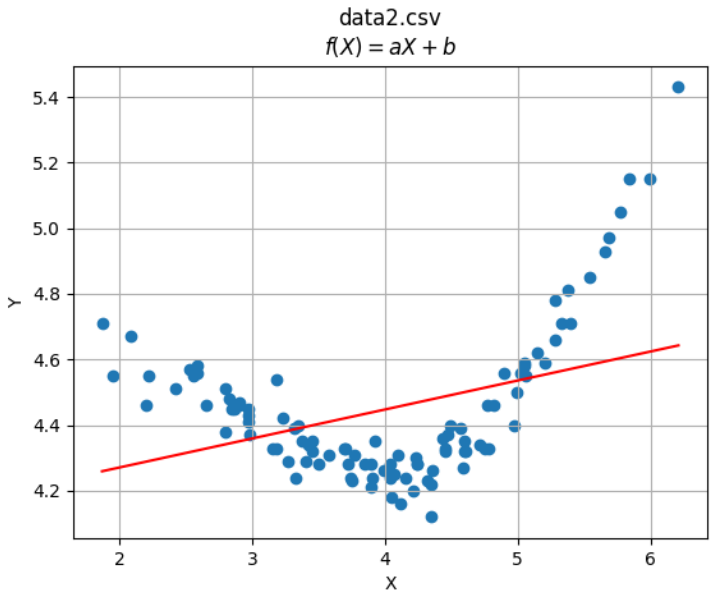
Otrzymana p-wartość: 0,5009497

Nie udało się odrzucić hipotezy.

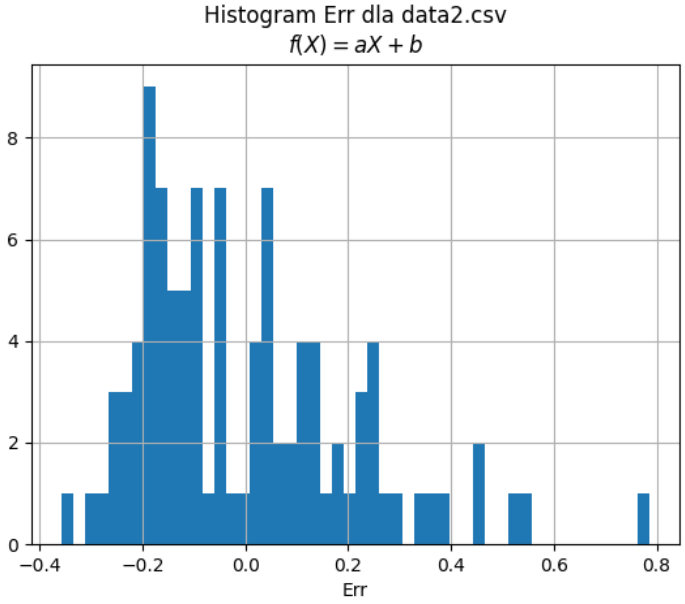
Tabela 4: Wyznaczone wartości parametrów dla modelu f(X) = aX + b i zestawu danych data2.csv

Parametr	Wartość
a	0,08837797436001704
b	4,09401704151636
średni błąd kwadratowy	0,044317955472194505
największa wartość odchylenia	0,7871557377079341
współczynnik R^2	0,1505413732999593

Wykres 4: Wykres przedstawiający modelowaną funkcję f(X) = aX + b na tle punktów z zestawu danych data2.csv



Histogram 4: Histogram odchyleń wartości funkcji f(X) = aX + b od danych z zestawu danych data2.csv



Test hipotezy statystycznej dla $f(X)=aX+b$ i zestawu danych data2.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

Otrzymana p-wartość: 0,0010321

Hipoteza odrzucona. Nie wydaje się by błędy miały rozkład normalny.

Ocena przydatności

Biorąc pod uwagę współczynniki R^2 oraz testy zgodności χ^2 Pearsona, dopasowanie modelu $f(X)=aX+b$ do zestawu danych data1.csv można uznać za sukces (R^2 powyżej 85% oraz p-wartość większa od ustalonego poziomu istotności). Wizualnie również dopasowanie wygląda poprawnie. Inaczej jest w przypadku zestawu danych data2.csv, gdzie współczynnik R^2 posiada wartość około 15%, a test zgodności χ^2 Pearsona zaprzeczył by błędy miały rozkład normalny. Wizualnie również regresja odbiega od krotek.

Model $f(X)=aX^2+b\sin(X)+c$

W celu przygotowania danych do zastosowania prostej regresji liniowej, sztucznie wprowadziłem $X_1=X^2$ oraz $X_2=\sin(X)$ i rozwiązywałem problem dla $g(X_1,X_2)=c+a\,X_1+b\,X_2$.
Utworzyłem macierz X:

$$X=\begin{bmatrix}1 & (X_1)_1 & (X_2)_1\\ 1 & (X_1)_2 & (X_2)_2\\ \vdots & \vdots & \vdots\\ 1 & (X_1)_n & (X_2)_n\end{bmatrix}$$

oraz macierz kolumnową A:

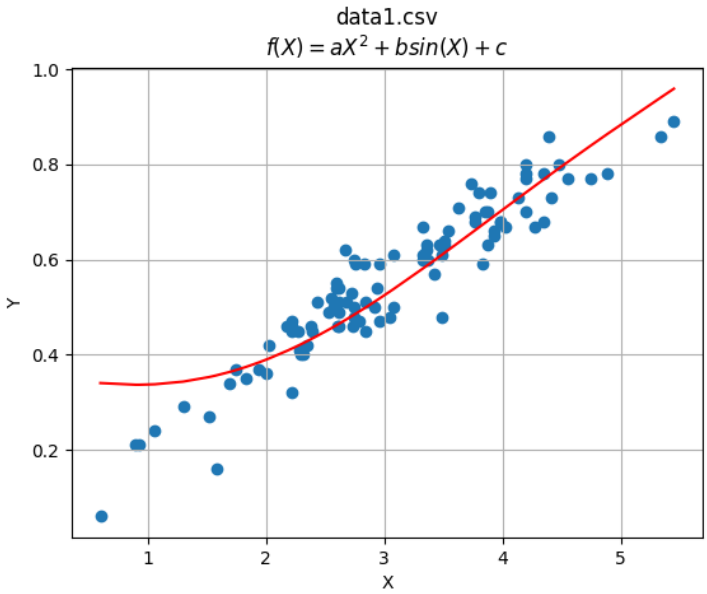
$$A=\begin{bmatrix}c\\ a\\ b\end{bmatrix} .$$

Wykorzystując udostępniony wzór w materiałach do tego zadania: $A=(X^T X)^{-1}X^T Y$ oraz weryfikując, że dla obydwu zestawów danych data1.csv i data2.csv macierz $X^T X$ jest odwracalna, otrzymałem parametry a, b oraz c.

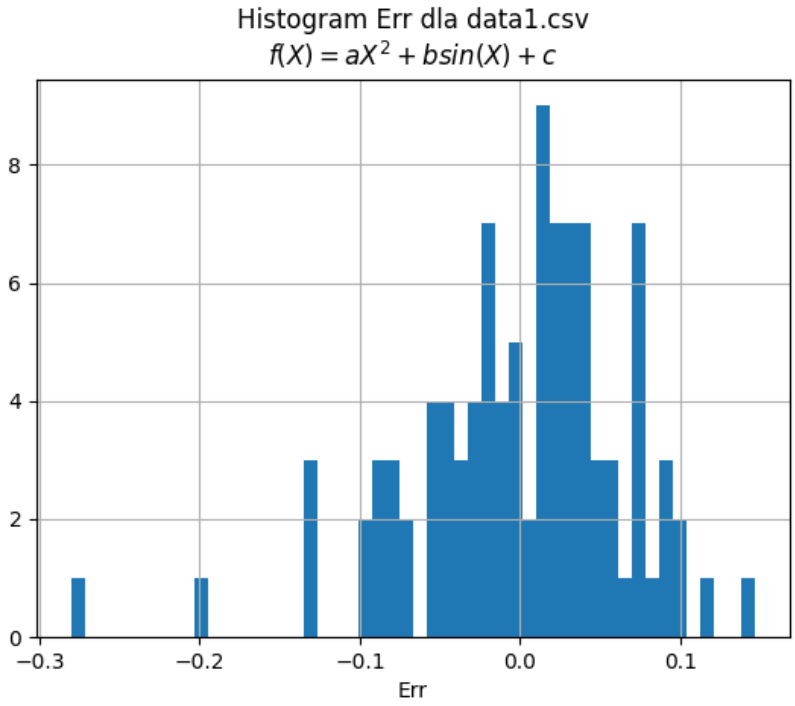
Tabela 5: Wyznaczone wartości parametrów dla modelu $f(X) = aX^2 + b\sin(X) + c$ i zestawu danych data1.csv

Parametr	**Wartość**
a	0,018666291388895973
b	-0,05467282974021806
c	0,36462103685000674
średni błąd kwadratowy	0,0043784975733911794
największa wartość odchylenia	0,2804702999379869
współczynnik R^2	0,8313627494457256

Wykres 5: Wykres przedstawiający modelowaną funkcję $f(X) = aX^2 + bsin(X) + c$ na tle punktów z zestawu danych data1.csv



Histogram 5: Histogram odchyleń wartości funkcji $f(X) = aX^2 + bsin(X) + c$ od danych z zestawu danych data1.csv



Test hipotezy statystycznej dla $f(X) = aX^2 + b\sin(X) + c$ i zestawu danych data1.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

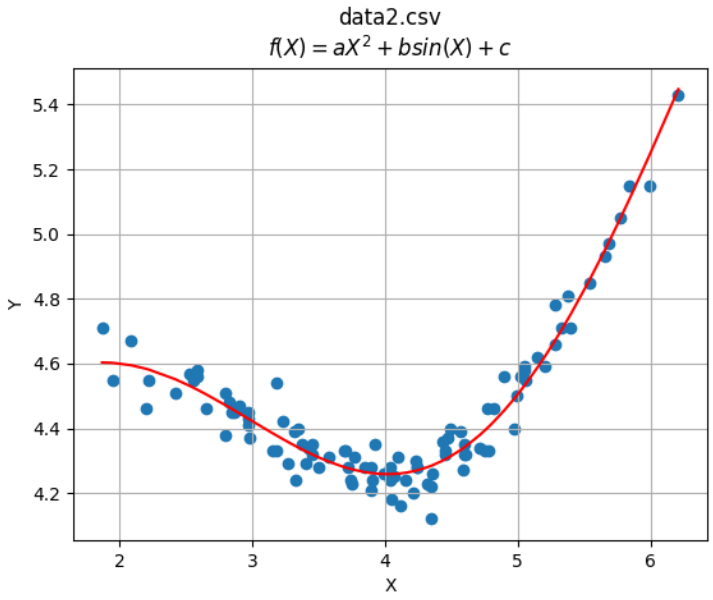
Otrzymana p-wartość: 0,0162181

Hipoteza odrzucona. Nie wydaje się by błędy miały rozkład normalny.

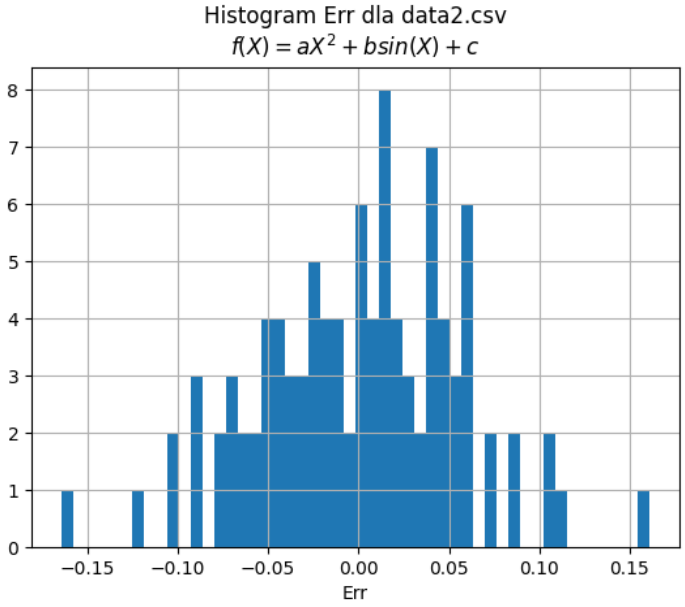
Tabela 6: Wyznaczone wartości parametrów dla modelu $f(X) = aX^2 + bsin(X) + c$ i zestawu danych data2.csv

Parametr	**Wartość**
a	0,03815351885214766
b	0,4799181398853723
c	4,011370092980355
średni błąd kwadratowy	0,0030385169504359016
największa wartość odchylenia	0,16458135974324595
współczynnik R^2	0,9417596229694875

Wykres 6: Wykres przedstawiający modelowaną funkcję $f(X) = aX^2 + bsin(X) + c$ na tle punktów z zestawu danych data2.csv



Histogram 6: Histogram odchyleń wartości funkcji $f(X) = aX^2 + bsin(X) + c$ od danych z zestawu danych data2.csv



Test hipotezy statystycznej dla $f(X)=aX^2+b\sin(X)+c$ i zestawu danych data2.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

Otrzymana p-wartość: 0,2442608

Nie udało się odrzucić hipotezy.

Ocena przydatności

Biorąc pod uwagę współczynniki R^2 oraz testy zgodności χ^2 Pearsona, dopasowanie modelu $f(X)=aX^2+b\sin(X)+c$ do zestawu danych data1.csv nie można uznać za pełny sukces (R^2 powyżej 83% lecz p-wartość mniejsza od ustalonego poziomu istotności). Wizualnie dopasowanie „mniej-więcej" wygląda poprawnie, lecz nie całkowicie. Inaczej jest w przypadku zestawu danych data2.csv, gdzie współczynnik R^2 posiada wartość powyżej 94%, a test zgodności χ^2 Pearsona nie zaprzeczył by błędy nie miały rozkładu normalnego. Wizualnie również regresja wygląda na dopasowaną.

Model $f(X_1, X_2)=aX_1+bX_2+c$

W celu przygotowania danych do zastosowania regresji liniowej wielokolumnowej, utworzyłem macierz X:

$$X=\begin{bmatrix} 1 & (X_1)_1 & (X_2)_1 \\ 1 & (X_1)_2 & (X_2)_2 \\ \vdots & \vdots & \vdots \\ 1 & (X_1)_n & (X_2)_n \end{bmatrix}$$

oraz macierz kolumnową A:

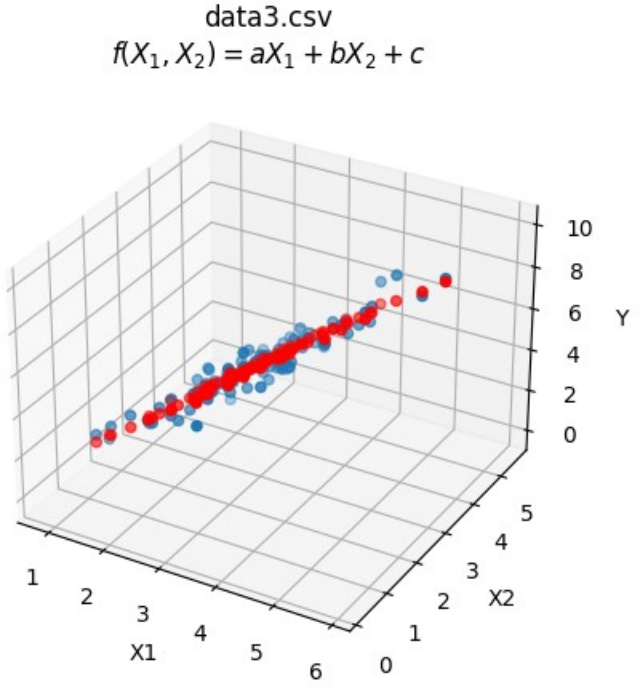
$$A=\begin{bmatrix} c \\ a \\ b \end{bmatrix} .$$

Wykorzystując udostępniony wzór w materiałach do tego zadania: $A=(X^T X)^{-1} X^T Y$ oraz weryfikując, że dla obydwu zestawów danych data3.csv i data4.csv macierz $X^T X$ jest odwracalna, otrzymałem parametry a, b oraz c.

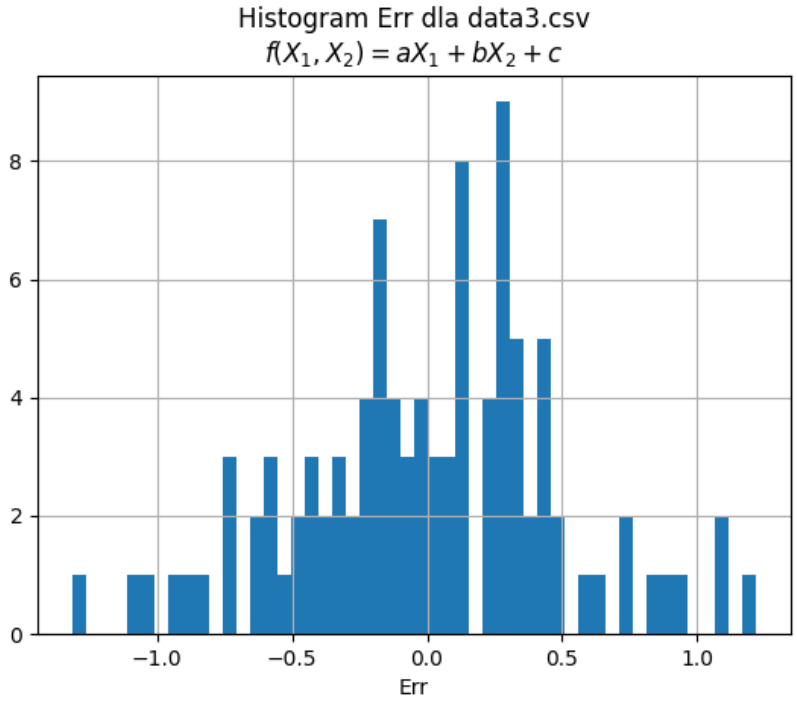
Tabela 7: Wyznaczone wartości parametrów dla modelu $f(X) = aX_1 + bX_2 + c$ i zestawu danych data3.csv

Parametr	**Wartość**
a	1,9571884273957556
b	-0,470592101592402
c	0,03002528405025351
średni błąd kwadratowy	0,23705954337678273
największa wartość odchylenia	1,3161075686191461
współczynnik R^2	0,9368753384171704

Wykres 7: Wykres przedstawiający modelowaną funkcję $f(X) = aX_1 + bX_2 + c$ na tle punktów z zestawu danych data3.csv



Histogram 7: Histogram odchyleń wartości funkcji f(X) = $aX_1 + bX_2 + c$ od danych z zestawu danych data3.csv



Test hipotezy statystycznej dla $f(X_1, X_2) = aX_1 + bX_2 + c$ i zestawu danych data3.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

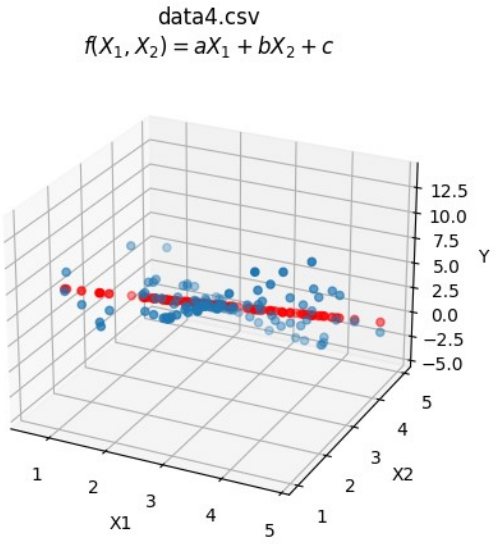
Otrzymana p-wartość: 0,3027164

Nie udało się odrzucić hipotezy.

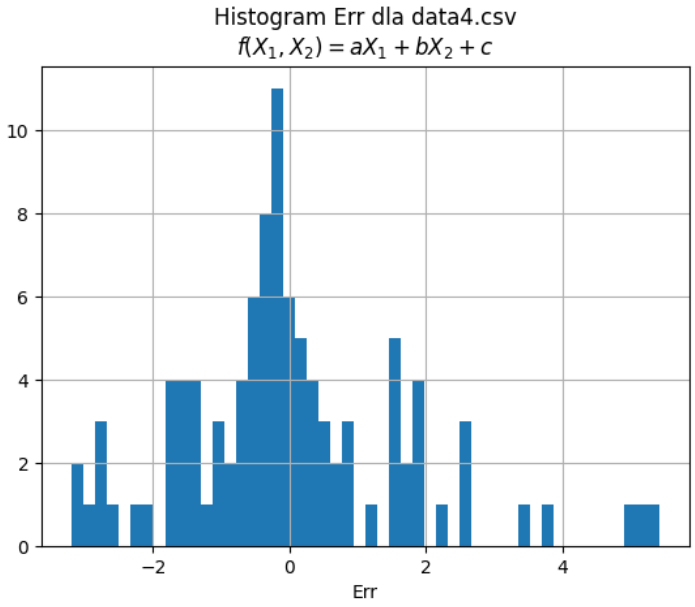
Tabela 8: Wyznaczone wartości parametrów dla modelu f(X) = $aX_1 + bX_2 + c$ i zestawu danych data4.csv

Parametr	**Wartość**
a	0,5566436121865908
b	-2,7966899080558405
c	10,045655157792132
średni błąd kwadratowy	2,773588109878039
największa wartość odchylenia	5,433347177115195
współczynnik R^2	0,8010570202376632

Wykres 8: Wykres przedstawiający modelowaną funkcję f(X) = $aX_1 + bX_2 + c$ na tle punktów z zestawu danych data4.csv



Histogram 8: Histogram odchyleń wartości funkcji f(X) = $aX_1 + bX_2 + c$ od danych z zestawu danych data4.csv



Test hipotezy statystycznej dla $f(X_1, X_2) = aX_1 + bX_2 + c$ i zestawu danych data4.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

Otrzymana p-wartość: 0,0086767

Hipoteza odrzucona. Nie wydaje się by błędy miały rozkład normalny.

Ocena przydatności

Biorąc pod uwagę współczynniki R^2 oraz testy zgodności χ^2 Pearsona, dopasowanie modelu $f(X_1, X_2) = aX_1 + bX_2 + c$ do zestawu danych data3.csv można uznać za sukces (R^2 powyżej 93%, p-wartość większa od ustalonego poziomu istotności). Wizualnie dopasowanie wygląda poprawnie. Inaczej jest w przypadku zestawu danych data4.csv, gdzie współczynnik R^2 posiada wartość powyżej 80%, ale test zgodności χ^2 Pearsona zaprzeczył by błędy miały rozkład normalny. Wizualnie regresja wygląda na częściowo dopasowaną, ale nie idealną.

Model $f(X_1, X_2) = a X_1^2 + bX_1X_2 + c X_2^2 + dX_1 + eX_2 + f$

W celu przygotowania danych do zastosowania regresji liniowej wielokolumnowej, sztucznie wprowadziłem:

$$X_3 = X_1^2 \quad , \quad X_4 = X_1 X_2 \quad , \quad X_5 = X_2^2$$

oraz rozwiązałem problem dla $g(X_1, X_2, X_3, X_4, X_5) = f + aX_3 + bX_4 + cX_5 + dX_1 + eX_2$. Utworzyłem macierz X:

$$X = \begin{bmatrix} 1 & (X_3)_1 & (X_4)_1 & (X_5)_1 & (X_1)_1 & (X_2)_1 \\ 1 & (X_3)_2 & (X_4)_2 & (X_5)_2 & (X_1)_2 & (X_2)_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (X_3)_n & (X_4)_n & (X_5)_n & (X_1)_n & (X_2)_n \end{bmatrix}$$

oraz macierz kolumnową A:

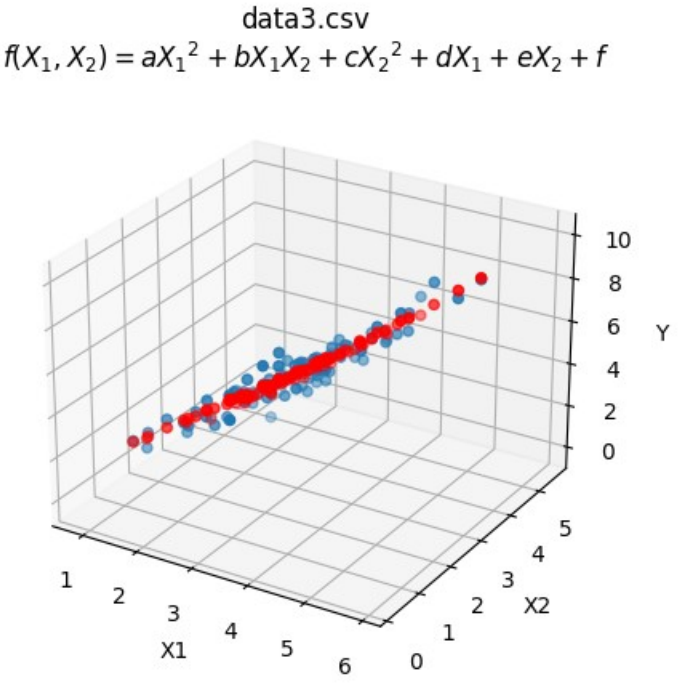
$$A = \begin{bmatrix} f \\ a \\ b \\ c \\ d \\ e \end{bmatrix} .$$

Wykorzystując udostępniony wzór w materiałach do tego zadania: $A = (X^T X)^{-1} X^T Y$ oraz weryfikując, że dla obydwu zestawów danych data3.csv i data4.csv macierz $X^T X$ jest odwracalna, otrzymałem parametry a, b oraz c.

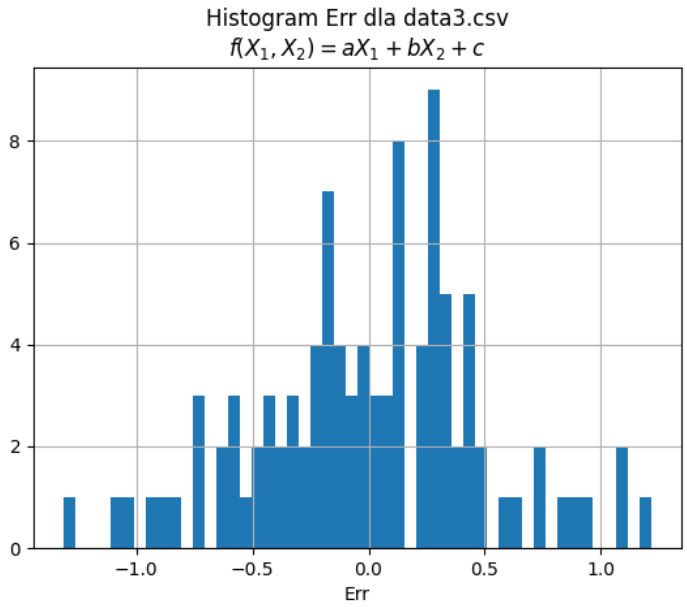
Tabela 9: Wyznaczone wartości parametrów dla modelu $f(X) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ i zestawu danych data3.csv

Parametr	Wartość
a	0,019747330459822632
b	0,07635587234881534
c	0,005775786373769687
d	1,5928374571903245
e	-0,7162798873498508
f	0,904659387516312
średni błąd kwadratowy	0,22936381845980702
największa wartość odchylenia	1,2242279560771139
współczynnik R^2	0,9389245705387665

Wykres 9: Wykres przedstawiający modelowaną funkcję $f(X) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ na tle punktów z zestawu danych data3.csv



Histogram 9: Histogram odchyleń wartości funkcji $f(X) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ od danych z zestawu danych data3.csv



Test hipotezy statystycznej dla $f(X_1, X_2) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ i zestawu danych data3.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

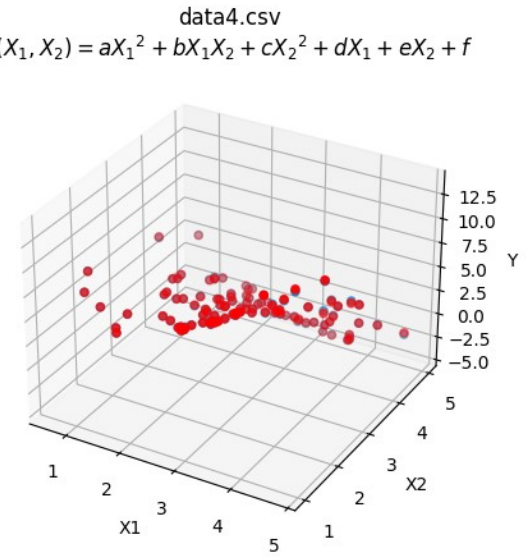
Otrzymana p-wartość: 0,2935389

Nie udało się odrzucić hipotezy.

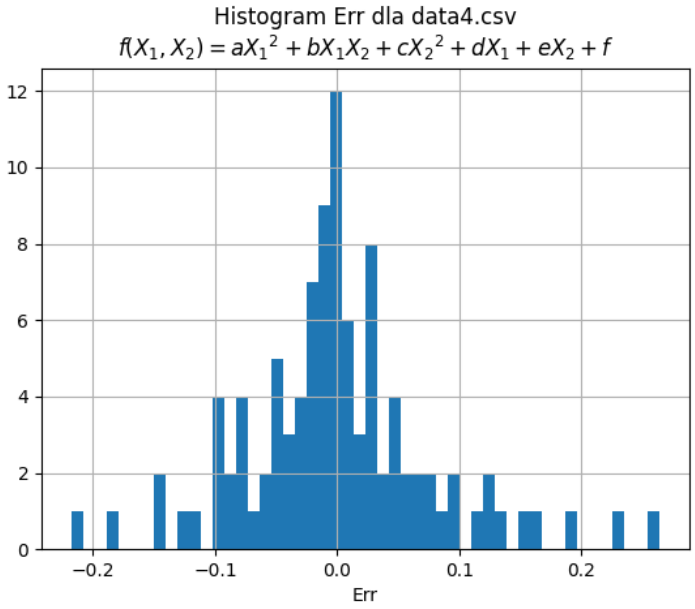
Tabela 10: Wyznaczone wartości parametrów dla modelu $f(X) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ i zestawu danych data4.csv

Parametr	**Wartość**
a	1,0009434346442514
b	-0,9948591701872805
c	-0,5128411410396918
d	-2,016271175898374
e	3,056891292809773
f	4,9662168934353925
średni błąd kwadratowy	0,006057403145268767
największa wartość odchylenia	0,2643063244786532
współczynnik R^2	0,9995655166579891

Wykres 10: Wykres przedstawiający modelowaną funkcję $f(X) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ na tle punktów z zestawu danych data4.csv



Histogram 10: Histogram odchyleń wartości funkcji $f(X) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ od danych z zestawu danych data4.csv



Test hipotezy statystycznej dla $f(X_1, X_2) = aX_1^2 + bX_1X_2 + cX_2^2 + dX_1 + eX_2 + f$ i zestawu danych data4.csv (test zgodności χ^2 Pearsona):

Hipoteza: Błędy mają rozkład normalny.

Poziom istotności: 0,05

Otrzymana p-wartość: 0,0396639

Hipoteza odrzucona. Nie wydaje się by błędy miały rozkład normalny.

Ocena przydatności

Biorąc pod uwagę współczynniki R^2 oraz testy zgodności χ^2 Pearsona, dopasowanie modelu $f(X_1, X_2) = aX_1 + bX_2 + c$ do zestawu danych data3.csv można uznać za sukces (R^2 powyżej 93%, p-wartość większa od ustalonego poziomu istotności). Wizualnie dopasowanie wygląda poprawnie. Inaczej jest w przypadku zestawu danych data4.csv, gdzie współczynnik R^2 posiada wartość powyżej 99,9%, ale test zgodności χ^2 Pearsona zaprzeczył by błędy miały rozkład normalny. Wizualnie regresja wygląda na bardzo dobrze.