**Uniwersytet Wrocławski**
**Wydział Matematyki i Informatyki**
**Instytut Matematyczny**
*specjalność: Analiza danych*

***Aneta Ewelina Przydróżna***

# Podstawowe techniki analizy statystycznej na przykładzie danych dotyczących emisji CO2

Praca licencjacka
napisana pod kierunkiem
dr hab. Macieja Paluszyńskiego

Wrocław 2020

# Spis treści

# 1 Wstęp

W pracy przedstawiono podstawowe techniki analizy statystycznej dużych zbiorów danych na przykładzie emisji $CO_2$ w tonach na jednego mieszkańca poszczególnych państw, w latach 1960-2014. Dzięki analizie statystycznej możemy opisać dokładnie wszystkie interesujące nas cechy na wiele sposobów. Począwszy od średniej arytmetycznej, która jedną wartością określa ogólną charakterystykę danego zbioru, przez regresję, która go wizualizuje, kończąc na testowaniu hipotez, określających słuszność naszych przypuszczeń. Jest to pomocne w późniejszych badaniach i może dać solidne podstawy w podejmowaniu decyzji[1].

W drugim rozdziale przedstawiono wszystkie potrzebne do badań fakty teoretyczne oraz definicje. Będzie to między innymi regresja i różne jej modele, a także sposoby ich dobierania, i porównywania. Ponadto dowiemy się na czym polega korelacja i co wyznaczają jej współczynniki.

W trzecim rozdziale przedstawione metody zastosowane zostały do analizy danych dotyczących emisji $CO_2$. Analiza ta pozwala wyróżnić szczególne przypadki odbiegające od powszechnych schematów. Dotyczy to w szczególności Chin oraz Islandii. Pozwala również na znajdowanie związków między danym państwami czy ich porównywanie.

Do tego celu wykorzystano język programowania Python w wersji 3.6.3, a także biblioteki *NumPy*, *Matplotlib*, *Pandas* oraz *scikit-learn*, które zawierają implementację narzędzi statystycznych wykorzystywanych w pracy. Więcej informacji na ten temat znajdzie się w rozdziale czwartym.

## 2 Teoretyczne podstawy stosowanych metod

Dane, na których opiera się ta praca, pochodzą ze zbiorów organizacji *The World Bank* [2].

| Państwo | 1960 | 1961 | $\cdots$ | 2014 |
|---|---|---|---|---|
| Afganistan | 0.046 | 0.054 | | 0.294 |
| Albania | 1.258 | 1.374 | | 1.978 |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| Polska | 6.740 | 6.923 | $\vdots$ | 7.517 |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| Zambia | 1.034 | 1.166 | | 0.292 |

Jak widać na powyższym fragmencie tabeli danych, tabela składa się z kolumny z nazwami wybranych państw świata, a także całych kontynentów (w kolejności alfabetycznej), oraz pięćdziesięciu pięciu kolumn ze średnimi wartościami emisji $CO_2$ na jednego mieszkańca. Można zatem przyjąć, że zmienna przedstawiająca zbiór $\{1960, 1961, \ldots, 2014\}$ jest zmienną dyskretną, nazwijmy ją $X$. Natomiast zbiór wartości, odpowiadający emisji $CO_2$ jest zmienną ciągłą $Y$ o wartościach rzeczywistych nieujemnych. Dzięki temu możemy na różne sposoby analizować dane, znajdować określone cechy lub porównywać wybrane państwa. Ponadto możemy szacować na ile dane państwo wpasowuje się w charakterystykę kontynentu, na którym się znajduje.

### 2.1 Regresja i niektóre jej modele

Pierwszą metodą analizy statystycznej będzie regresja, która pozwala opisać wzajemne powiązania kilku zmiennych za pomocą funkcji. Stosując ją, możemy przedstawić dane w matematyczny sposób, który pozwoli na dokładniejszą analizę cech. Możemy także próbować przewidzieć wartości zmiennej zależnej, w naszym przypadku zmiennej $Y$, używając znanych nam wartości zmiennych niezależnych, tutaj: $X$. Aby to zrobić, konstruujemy model, czyli funkcję, która w przestrzeni dwuwymiarowej może przybrać postać linii lub krzywej zadanej bardziej złożonym wzorem. W przestrzeni wielowymiarowej mówimy o hiperprzestrzeni. Celem jest najlepsze dopasowanie do danych, na przykład tak, aby zminimalizować różnicę kwadratów pomiędzy odległościami punktów danych, a wybraną krzywą. Gdy mamy już odpowiednią funkcję, podstawiamy do niej predyktory, czyli kolejne wartości $X$, aby uzyskać prognozowaną wartość zmiennej odpowiedzi, $Y$. Podsumowując, na początek dobieramy model, a następnie korzystając z odpowiednich narzędzi szacujemy wartości parametrów tego modelu.

**Definicja 1.** Ogólna postać modelu wygląda następująco:

$$Y = f(X, \beta) + \varepsilon,$$

gdzie $Y$ to zmienna zależna, $f(X, \beta)$ to funkcja regresji, której postać powinna odpowiadać charakterowi danych. Pierwszą wartością funkcji jest wektor lub macierz zmiennych niezależnych, a drugą jest wektor parametrów, od których zależą szczególne własności funkcji $f$. Parametry te powinniśmy dobrać tak, aby jak najlepiej pasowały do danych. Na przykład, jeżeli uznamy, że właściwym modelem dla danych jest dwuwymiarowy model liniowy, to wektor $\beta$ składa się z dwóch współczynników: $Y = \beta_0 + X\beta_1$. Dobrane wartości $\hat{\beta}$ nazywamy estymatorami $\beta$, a dobraną funkcję $\hat{Y}$ estymatorem zmiennej zależnej. Ostatnim składnikiem wzoru jest $\varepsilon$, czyli zmienna losowa (lub szum losowy), która mierzy odstępstwo konkretnych danych od modelu. Wartością oczekiwaną szumu jest 0.

Istnieje wiele metod regresji. Najczęściej stosowana jest regresja liniowa, używana do zmiennych o charakterze liniowym. Jej model wygląda następująco:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{in}\beta_n + \varepsilon_i \qquad i = 1, \ldots, p \tag{1}$$

Jest to ogólna postać modelu liniowego, gdzie $y_i$ to $i$-te przedstawienie $Y$ za pomocą $i$-tych wyrazów zmiennych objaśniających $X_1, \ldots, X_n$. Podane we wzorze $\beta_0$ to wyraz wolny, będący punktem początkowym, natomiast pozostałe współczynniki są powiązane z odpowiednimi zmiennymi niezależnymi. Błędy losowe $\varepsilon_i$ z rozkładu $N(0, \sigma^2)$, są niezależnymi zmiennymi z wartością oczekiwaną $\mathbb{E}(\varepsilon_i|X_i) = 0$, oraz wariancją $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Powyższy model nazywamy modelem regresji wielorakiej. Możemy go przedstawić w najprostszej formie otrzymując:

$$Y = \beta_0 + X\beta_1 + \varepsilon \tag{2}$$

Używając funkcji liniowych jako estymatorów, trzeba liczyć się z ich małą elastycznością. Zatem przy różnych typach danych możemy mieć problem z odpowiednim dopasowaniem.

Regresja liniowa może jednak być przedstawiona w wersji z dodatkowymi przekształceniami, na przykład:

$$Y = \beta_0 + \sqrt{X}\beta_1 + \log X\beta_2 + X^3\beta_3 + \varepsilon$$

lub w wersji z interakcjami, takimi jak:

$$Y = \beta_0 + X_2X_3\beta_1 + X_2\beta_2 + X_3\beta_n + \varepsilon$$

Mimo że nie są to funkcje liniowe, w łatwy sposób możemy je zlinearyzować odpowiednimi przekształceniami, tworząc liniową kombinację dzięki temu, że współczynniki $\beta$ zachowują liniowość.

Często spotykanym przypadkiem regresji liniowej, dopuszczającym inną niż prosta linię jest regresja wielomianowa. Poprzez podnoszenie zmiennej zależnej do wyższych potęg, pozwala ona na dokładniejsze dopasowanie się do zależności krzywoliniowej bez użycia bardzo skomplikowanych algorytmów. Poniżej znajduje się przykładowy wzór określający model wielomianowy:

$$Y = \beta_0 + X\beta_1 + X^2\beta_2 + X^3\beta_3$$

Przy regresji liniowej warto wspomnieć o jej specjalnym przypadku, gdzie zmienną objaśniającą $X$ jest zmienna czasowa $t$, a zmienna $Y$ jest monotoniczna. Opisywane zjawisko to tendencja rozwojowa (trend). Możemy rozróżnić trzy rodzaje trendów: rosnący, malejący lub boczny, który nie cechuje się wyraźnym wzrostem lub spadkiem wartości zmiennej. Aby stworzyć model trendu, należy dobrać odpowiednią monotoniczną funkcję zależną od czasu $t$. Funkcja trendu może przybierać postać na przykład funkcji liniowej czy logarytmicznej.

Czasem jednak pewne przekształcenia powodują, że modelu nie można wyrazić w postaci kombinacji liniowej wyrazów. Jest to regresja nieliniowa. Jej plusem jest większa elastyczność dopasowania. Oto przykładowa forma modelu, który nie sprowadza się do kombinacji liniowej.

$$Y = \beta_0 + \frac{\beta_1 X}{X - \beta_2} + \varepsilon$$

Kolejną popularną metodą, o której warto wspomnieć jest regresja logistyczna (odpowiednia do danych binarnych, na przykład 0/1, tak/nie lub sukces/porażka). Jest to uogólniony model liniowy, który wykorzystuje funkcję logit do przekształca prawdopodobieństwo zajścia danej sytuacji.

**Definicja 2.** Szansa to stosunek sukcesu ($p$) do porażki ($1 - p$). Zadana jest wzorem $\frac{p_i}{1-p_i} = e^{\beta_0} e^{\beta_i x_{ik}}$, gdzie $x_{ik}$ to $i$-ty wyraz $k$-tej zmiennej niezależnej, a $\beta_i$ to $i$-ty współczynnik regresji.

Funkcja logit nakłada logarytm na szansę, dając znaną nam już funkcję:

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}.$$

Korzystając z powyższych równości możemy wyznaczyć estymator prawdopodobieństwa zajścia sukcesu:

$$\hat{p_i} = \frac{e^{logit(p_i)}}{1 + e^{logit(p_i)}} = \frac{1}{1 + e^{-logit(p_i)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni})}}.$$

Istnieje jeszcze wiele metod regresji, a każda z nich sprawdza się lepiej w innych przypadkach, takich jak wielowymiarowość zmiennych niezależnych, duża powtarzalność próbek czy niewielka liczba danych. Wybór modelu nie jest oczywisty, aczkolwiek możemy kierować się głównymi przeznaczeniami każdego z nich i sprawdzić który schemat najlepiej pasuje do naszego problemu. Dla przykładu, najlepszym sposobem podziału cechy ze względu na płeć będzie regresja logistyczna. Gdy mamy mocno wzajemnie powiązane predyktory, używamy regresji nazywanej grzbietową. Kiedy podczas badania wpływu czasu na ilość wyprodukowanych przedmiotów zauważamy niemalże monotoniczne zależności najprościej będzie użyć regresji liniowej. Gdy chcemy przewidzieć czy użytownik przedłuży subskrypcje sięgniemy po model lasów losowych. Jak widać, każdy problem można rozwiązać, po dokładniejszym przyjrzeniu się danym, a w przypadku dalszych wątpliwości, warto jest użyć metod porównawczych i na ich podstawie wybrać lepiej dopasowany model.

## 2.2 Metoda najmniejszych kwadratów

Błąd oszacowania to różnica między linią regresji a prawdziwymi wartościami $Y$. Procedura nazywana metodą najmniejszych kwadratów polega na zminimalizowaniu sumy kwadratów tej różnicy. Oznacza to, że nie istnieje inna funkcja dla danej regresji, która dałaby mniejszą wartość błędu. Powróćmy do wzoru (1). Zobaczymy, w jaki sposób można znaleźć model regresji wielorakiej najlepiej dopasowany do danych. Załóżmy, że wektor współczynników oraz szum losowy $\varepsilon$ mają długość $n+1$ oraz, że mamy $n+1$ zmiennych objaśniających długości $p$, z czego pierwsza jest złożona z samych jedynek (odpowiada ona wyrazowi wolnemu $\beta_0$). Zmienne te tworzą macierz $p \times (n+1)$, nazwaną $X$. Jej kolumny są liniowo niezależne. Można teraz przedstawić prawdziwe wartości $Y$ jako $X\beta+\varepsilon$, a dopasowany model regresji jako $\hat{Y} = X\hat{\beta}$. Prawdziwe wartości wektora parametrów $\beta$ są nieznane. Aby odnaleźć optymalny dla naszego modelu estymator $\beta$, czyli wektor $\hat{\beta}$ stosujemy metodę najmniejszych kwadratów. Minimalizując sumę kwadratów reszt $\sum(y_i - x_i^T\hat{\beta})^2$ otrzymujemy minimum globalne w podanej formule:

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

gdzie $X^T$ to transponowana macierz $X$. Mając wyestymowane współczynniki $\hat{\beta}$, możemy przewidywać wartości $\hat{Y}$. Powyższe rozumowanie bardzo się upraszcza, gdy mamy tylko jedną zmienną niezależną o indeksach $i = 1, 2, \ldots, p$. Korzystając ze wzoru (2) i faktu, że $\varepsilon_i \sim N(0, \sigma^2)$ oraz $\mathbb{E}(\varepsilon_i|x_i) = 0$, możemy stwierdzić, że $\mathbb{E}(y_i|x_i) = \beta_0 + x_i\beta_1$ oraz $\mathrm{Var}(y_i|x_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$. Tworząc zmienną $\hat{Y} = \hat{\beta}_0 + X\hat{\beta}_1$, i przyrównując ją do prawdziwych wartości $Y$, otrzymujemy wartości resztowe $\varepsilon_i = y_i - \hat{y}_i$. Minimalizujemy sumę kwadratów odległości między prawdziwymi danymi a dopasowywaną przez nas krzywą.

Estymatory $\beta_0$ i $\beta_1$ wyglądają następująco:

$$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

gdzie $\overline{x}$ jest średnią zmiennej objaśniającej, a $\overline{y}$ jest średnią zmiennej objaśnianej. Podstawiamy estymatory do wzoru $\hat{Y} = \hat{\beta}_0 + X\hat{\beta}_1$. Korzystając z tego oszacowania, możemy przewidywać wartości dla kolejnych predyktorów.

Metoda najmniejszych kwadratów nie jest odporna na wartości odstające (nietypowe, rzadko występujące), ponieważ kwadrat ich odległości mocno wpływa na kierunek prostej w regresji liniowej. Powoduje to gorsze dopasowanie do pozostałych (nieodstających) wartości. W związku z tym czasem usuwa się nietypowe przypadki z bazy, by nie wpływały negatywnie na model.

### 2.3 Sposoby porównywania modeli

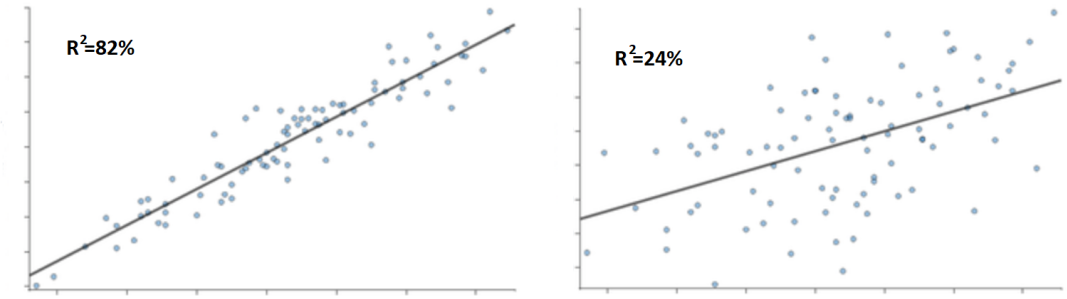
Kiedy już dopasujemy model do danych, warto by było sprawdzić, czy był to dobry wybór, czy też inny rodzaj regresji lepiej estymowałby $Y$. Jest wiele sposobów sprawdzania jakości dopasowania i porównywania modeli. Są to najczęściej miary dające ocenę liczbową, tego jak odchylenia punktów pomiarowych mają się do przewidywanej powierzchni regresji.

**Definicja 3.** Współczynnikiem determinacji nazywamy

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \qquad R^2 \in [0, 1]$$

gdzie $y_i$ to prawdziwa wartość zmiennej objaśnianej w chwili $i$, $\overline{y}$ jest średnią arytmetyczną empirycznych wartości zmiennych zależnych, a $\hat{y}_i$ określa wartość estymatora zmiennej objaśnianej w chwili $i$. $R^2$ często wyrażany jest w procentach.

Ta metoda mówi nam jaki procent $Y$ został wyjaśniony regresją względem $X$. Zależy nam na jak największej wartości $R^2$, ponieważ wtedy punkty danch są mniej rozproszone wokół dopasowanej lini regresji. Zazwyczaj im większy jest współczynnik determinacji w tym samym zestawie danych, tym mniejsze są wartości resztowe, czyli odstępy prawdziwych wartości zmiennej zależnej od modelu. Mimo krzywoliniowości modelu wielomianowego, współczynnik determinacji może być również stosowany do oceny jego jakości dopasowania. Natomiast do regresji nieliniowej nie nada się, ponieważ nie zawsze należałby do zdefiniowanego przedziału, co więcej nie rozróżniałaby dobrych i złych modeli.




Rysunek 1: Zestawienie danych o różnych rozproszeniach wokół linii regresji

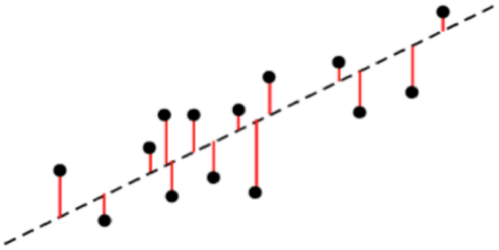
Powiązany ze współczynnikiem determinacji jest współcznnik zbieżności $\phi^2 = 1 - R^2$, który określa poziom niedopasowania zmiennej objaśnianej do modelu. Jej przedział to również $[0, 1]$, choć zazwyczaj wyrażana jest w procentach. Im $\phi^2$ jest bliżej zera, tym model powinien być lepiej dopasowany.

Zanim poznamy kolejną miarę jakości modeli, warto jest przyjrzeć się dwóm pojęciom : reszta i błąd statystyczny. To pierwsze oznacza różnicę między prawdziwą wartością obserwowaną a jej estymatorem. Natomiast błąd jest odchyleniem wartości obserwowanej od teoretycznej wartości, której nie jesteśmy w stanie obliczyć.

Najprostszą miarą dopasowania jest suma błędów $SE = \sum(\hat{y}_i - y_i)$, czyli suma reszt. Niestety często na jej podstawie możemy wyciągnąć błędne wnioski. Gdy prawdziwe wartości zmiennej obserwowanej są równomiernie położone po obu stronach funkcji regresji, neutralizują się, dając niską wartość $SE$. Nałożenie wartości bezwzględnej, choć wydaje się być pożyteczne, powoduje nieróżniczkowalność w punkcie 0, więc nie jest powszechnie stosowane. Inną metodą zaradzenia zerowania się odchyleń jest podniesienie różnic wartości $Y$ i $\hat{Y}$ do kwadratów. Minusem takiego podejścia jest ciągły wzrost wartości sumy kwadratów reszt, przy wzroście liczby obserwacji. Aby temu zaradzić dzielimy naszą sumę przez liczbę obserwacji, otrzymując *Mean Squared Errors*, czyli miarę odporną na wzrost liczby obserwacji. Wzór na nią jest następujący $MSE = \frac{1}{n}\sum(\hat{y}_i - y_i)^2$. Czasem nakładamy pierwiastek na $MSE$, tworząc $RMSE$ (*Root Mean Squared Error*).

Błąd standardowy regresji $S = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{(n-2)}}$ również ma nieco mylącą nazwę, ponieważ tak na prawdę jego wzór zawiera sumę kwadratów reszt a nie nieobserwowalnych błędów. Jego wzór jest bardzo podobny do wzoru na $RMSE$, lecz nie dzielimy kwadratów odchyleń przez liczbę obserwacji $n$, a przez liczbę stopni swobody $(n - p - 1)$, gdzie $p$ to liczba zmiennych objaśniających (w przypadku prostej regresji liniowej $p = 1$)[13].

Miara $S$ obowiązuje zarówno modele liniowe jak i nieliniowe, zatem przy porównywaniu ich warto sięgnąć po tę metodę. Tak jak w przypadku $MSE$ zależy nam na zminimalizowaniu błędu standardowego, zatem im mniejsza wartość $S$ w tym samym zestawie danych, tym bliżej znajdują się one lini regresji.



Rysunek 2: Odległości wartości resztowych od modelu w regresji liniowej

## 2.4 Korelacja i jej współczynniki

Korelacja określa współzależność zmiennych losowych i wyraża ją liczbą, nazywaną współczynnikiem korelacji. W tej pracy będziemy rozważać zmienne rzeczywiste $X$ i $Y$, występujące w zdarzeniach, określonych na przestrzeni probabilistycznej $(\Omega, \mathcal{F}, P)$. Istnieje wiele wzorów pozwalających na określenie stopnia współzależności. Jednym z częściej stosowanych do danych liczbowych jest współczynnik korelacji Pearsona. Stosuje się go do zależności między zmiennymi określanymi jako monotoniczne, w szczególności liniowe. Jednakże możemy go również zastosować w przypadku zależności kwadratowych, sześciennych i tym podobnych. Znormalizowany wynik przybiera wartość z przedziału [–1,1]. Zupełna korelacja ujemna zachodzi, gdy otrzymamy –1, co oznacza, że zmienne losowe mają przeciwne cechy. Gdy otrzymamy 0, uznajemy, że zmienne nie mają cech wspólnych. Natomiast, gdy wynikiem będzie 1, mamy zupełną korelację dodatnią, gdzie wzrost wartości jednej cechy oznacza proporcjonalny wzrost wartości drugiej. Siła związków korelacyjnych może być różnie interpretowana, lecz zawsze im niższa jest jej wartość bezwzględna, tym zależność jest mniejsza [11].

**Definicja 4** (Współczynnik korelacji Pearsona)**.** Niech $x_i$, dla $i = 1, 2, \cdots, n$, będą wartościami zmiennej $X$, a $y_i$ wartościami zmiennej $Y$. Średnie tych zmiennych będą oznaczone jako $\overline{x}$ oraz $\overline{y}$, wtedy współczynnik korelacji Pearsona jest wyrażony następującym wzorem:

$$r_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Dużą wadą stosowania powyższej miary jest jej wrażliwość na obserwacje odstające, które mają duży wpływ na wartość $r_{XY}$. Warto zatem sprawdzić czy nasze dane zawierają takie obserwacje i w miarę możliwości wykonać badania wyłączając je. Można też zastosować inną metodę, bazującą na współczynniku korelacji Pearsona. Jest to współczynnik korelacji rang Spearmana, stosowany do zmiennych o zależnościach monotonicznych, ale niekoniecznie liniowych lub do danych jakościowych. Nakłada on rangi na zmienne, normalizując obserwacje nietypowe. Rangi całych zmiennych zapisujemy w następujący sposób: $R_X$, $R_Y$. Natomiast rangi $i$-tych wyrazów tych zmiennych jako $R_{X_i}$ oraz $R_{Y_i}$. Poniżej znajduje się uproszczony wzór, który nie uwzględnia rang wiązanych.

**Definicja 5** (Współczynnik korelacji rang Spearmana )**.**

$$r_s = \frac{cov(R_X, R_Y)}{\sqrt{var(R_X)var(R_Y)}} = 1 - \frac{6 \cdot \sum_{i=1}^{n}(R_{x_i} - R_{y_i})^2}{n(n^2-1)}$$

Obie miary mają taki sam przedział i interpretację. Czasem jednak wykazują korelację bliską 0, mimo silnej zależności zmiennych. Na przykład w przypadku niemonotonicznego, a tym bardziej nieliniowego powiązania. Wówczas możemy zastosować inną metodę analizy.

**Definicja 6** (Niezależność statystyczna)**.** Zmienne $X$ i $Y$ są niezależne, gdy dla wszystkich liczb rzeczywistych $a$ i $b$ zachodzi równość

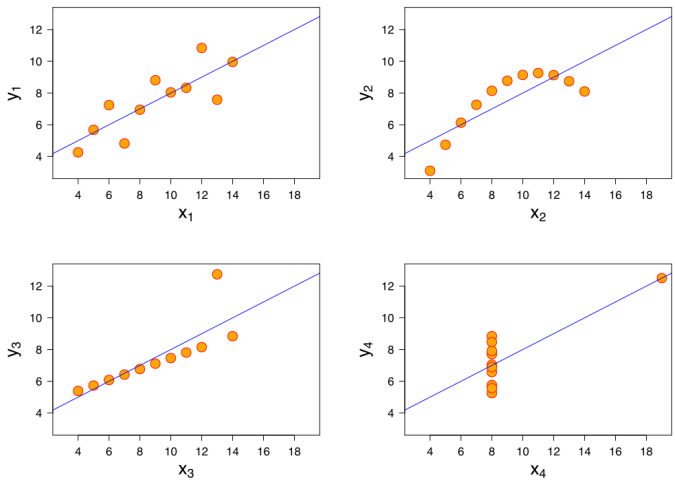
$$P(X \geqslant a)P(Y \geqslant b) = P(X \geqslant a \wedge Y \geqslant b)$$

**Definicja 7** (Zależność statystyczna)**.** Zmienne losowe $X$ i $Y$ są zależne, gdy nie są one niezależne, to znaczy, że dla pewnych liczb rzeczywistych $a$ i $b$ zachodzi

$$P(X \geqslant a)P(Y \geqslant b) \neq P(X \geqslant a \wedge Y \geqslant b)$$

Weźmy zmienne $X$ i $Y$, powiązane następującym wzorem: $Y = X^2$. Wykazują one silną zależność statystyczną na przedziale $X \in [-1, 1]$, gdy współczynniki korelacji są bardzo małe. Jest to przykład na to, że powinniśmy przyglądać się danym dokładnie, uwzględniając ich różne cechy. Wtedy nasza analiza statystyczna będzie solidną podstawą do ich oceny.

## 2.5 Kwartet Anscombe'a




Rysunek 3: Zestawienie różnych zbiorów danych o tych samych cechach statystycznych

Powyższy układ danych został stworzony przez Francisa Anscombe'a w celu ukazania znaczenia analizy danych, również wizualnie. Wszystkie cztery przypadki mają identyczne (do trzech miejsc po przecinku) podstawowe cechy statystyczne. Między innymi:

- Współczynnik korelacji równy 0.816
- Równanie regresji zadane wzorem $Y = 3 + 0.5X$
- Średnia arytmetyczna zmiennej $X$ równa 9
- Średnia arytmetyczna zmiennej $Y$ równa 7.5

Ten przykład pokazuje jak ważne jest wykonanie wyczerpujących badań. Każdy wynik testu lub statystyki jest cenną informacją, lecz nie zawsze daje pełny obraz cech danych. W kolejnym rozdziale tej pracy postaramy się uwzględnić wiele informacji, a także przyjrzeć się wykresom emisji $CO_2$ względem lat, by otrzymać rzetelne wyniki.

# 3 Zastosowanie przedstawionych technik analizy na zbiorze danych

W tym rozdziale spróbujemy porównać prostą regresję liniową oraz regresję wielomianową. Biorąc pod uwagę, że w latach 1960-2014 miało miejsce wiele wydarzeń polityczno-ekonomicznych, które znacząco wpłynęły na dynamikę zmian związanych z wytwarzaniem dwutlenku węgla, możemy przypuszczać, że lepszym wyborem będzie regresja wielomianowa, choć wpływ na emisję miało bardzo dużo czynników oprócz lat. W niewielu przypadkach wykres emisji $CO_2$ na jednego mieszkańca danego kraju był monotoniczny oraz liniowy, także również współczynnik korelacji Pearsona nie w każdej relacji między zmiennymi będzie odpowiednim narzędziem analizy.

Częstym elementem regresji jest początkowy podział danych na części. W wielu przypadkach jest to znaczne ułatwienie, wpływające korzystnie na dalsze prognozy. Dane możemy dzielić na kilka sposobów. Popularną metodą statystyczną jest walidacja krzyżowa. Na przykład $k$-krotna, która dzieli dany zbiór na $k$ równych podzbiorów. Te na których uczymy model, nazywamy treningowymi, natomiast te, na których sprawdzamy jakość estymacji modelu - testowymi. Liczba $k$ nie jest z góry określona i sami możemy ją dostosować według uznania. Następnie trenujemy model na $k-1$ podzbiorach i testujemy jakość dopasowania go na wyłączonym z uczenia podzbiorze. Gdy podzbiorów jest tyle ile elementów w zbiorze, mówimy o szczególnym przypadku walidacji $k$-krotnej, to znaczy o *leave-one-out*. Polega on na kolejnym wyłączaniu jednego elementu z próby i testowaniu modelu bez jego udziału. Możemy również losowo podzielić dane na dwa rozłączne podzbiory, gdzie testowy nie powinien zawierać więcej niż 50% zmiennych. Walidacja krzyżowa ma wiele zalet, między innymi zapobiega przetrenowaniu modelu, czyli nadgorliwemu dopasowaniu go do danych. Stanowi też wiarygodne źródło informacji na temat dopasowania, ponieważ jakość modelu jest oceniana na zbiorze, na którym nie był on wcześniej uczony. Co więcej stosując walidację przy odpowiednim doborze zbioru treningowego, otrzymamy dobry model bez konieczności pracy na całym zbiorze danych. To znacznie oszczędza czas, bez utraty jakości badań.

W tej pracy podzielimy zmienne $X$ (oraz odpowiadające im $Y$) między innymi w następujący sposób:

- Zbiór treningowy $\{1960, 1961, \ldots, 1999\}$
- Zbiór testowy $\{2000, 2001, \ldots, 2014\}$

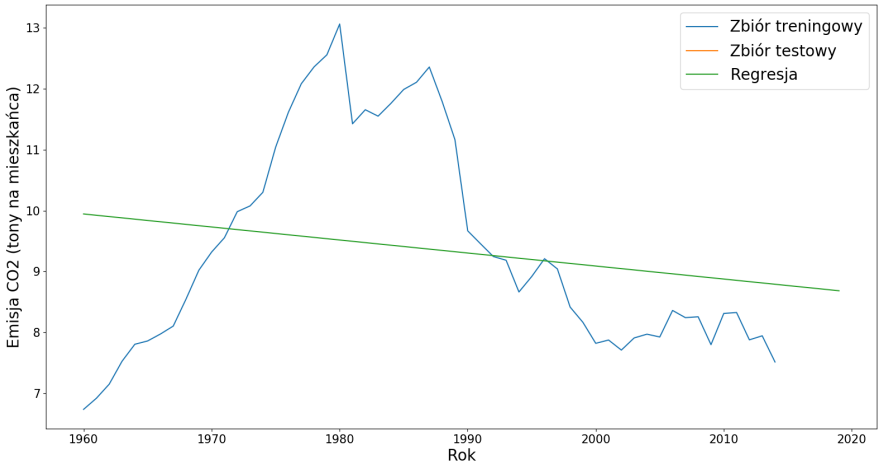
Zestawimy modele regresji wykonanej na podzbiorze danych, a także na wszystkich danych. Mimo, że zbiór danych kończy się na roku 2014, to model regresji przewiduje dane do bieżącego roku (2020).

### 3.1 Model liniowy

Jednym z wielu państw, których wykres emisji dwutlenku węgla nie jest monotoniczny jest Polska, gdzie maksymalną wartością $Y$ jest $y_{21} = 13.06$, natomiast wartości odpowiadające latom na krańcach przedziału to kolejno $y_1 = 6.74$ i $y_{55} = 7.52$. Można powiedzieć że wykres ten ma trend rosnący widoczny na przedziale 1960-1980. W kolejnych latach wartości zmiennej objaśnianej bardziej się wahają.

Linia prosta, jak się okaże, nie jest w stanie dobrze przedstawić tych zależności. Predykcja tym sposobem odbiega znacząco od prawdziwych wartości. W ostatnich latach, gdy wykres zaczyna się stabilizować, prosta regresji estymuje najlepiej. Bardzo prawdopodobne, że będzie się ona zbliżać do danych w kolejnych latach.

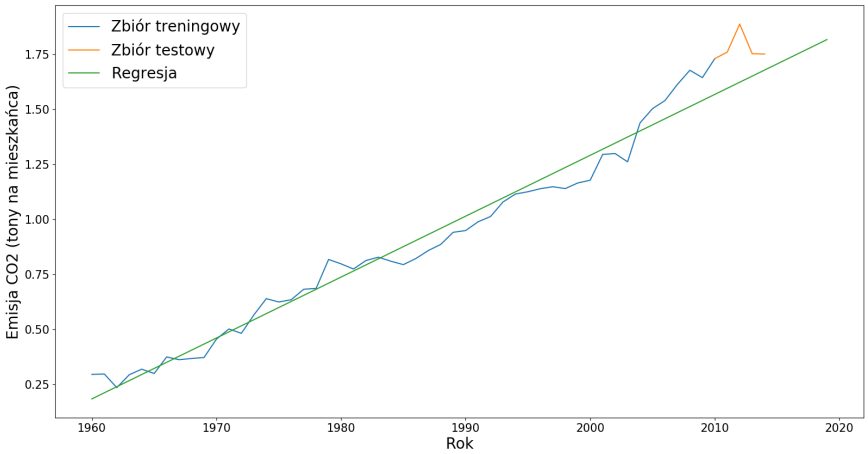
Linia regresji zaczyna się na wyskokości około 10, a kończy znacznie poniżej 9, przecinając średnią wartość $Y$ równą 9.37, oraz medianę równą 8.92. Zatem wykazuje zależność malejącą emisji $CO_2$ od lat, przewidując mniejszą jej wartość niż średnia na przestrzeni ostatnich dziesięcioleci.




Rysunek 4: Regresja liniowa dla Polski

Kolejny wykres przedstawia emisję $CO_2$ na jednego mieszkańca Maroka. W tym przypadku regresja liniowa sprawdza się znacznie lepiej niż w przypadku wykresu dla Polski. Jest to spowodowane bliskiemu do liniowego wzrostowi zmiennej $Y$ względem zmiennej $X$. Tym razem zbiorem treningowym nie jest cała krzywa, a jej fragment do roku 2010. Prawdziwa wartość $y_{55}$, która nie była włączona do tego zbioru, jest zbliżona do wartości jej estymatora, zatem można stwierdzić, że model dobrze przewiduje nieznane dla siebie dane.

Warto zwrócić uwagę jak małe są wartości $Y$ dla Maroka, mimo ciągłego wzrostu, w porównaniu do wartości tej zmiennej dla Polski.

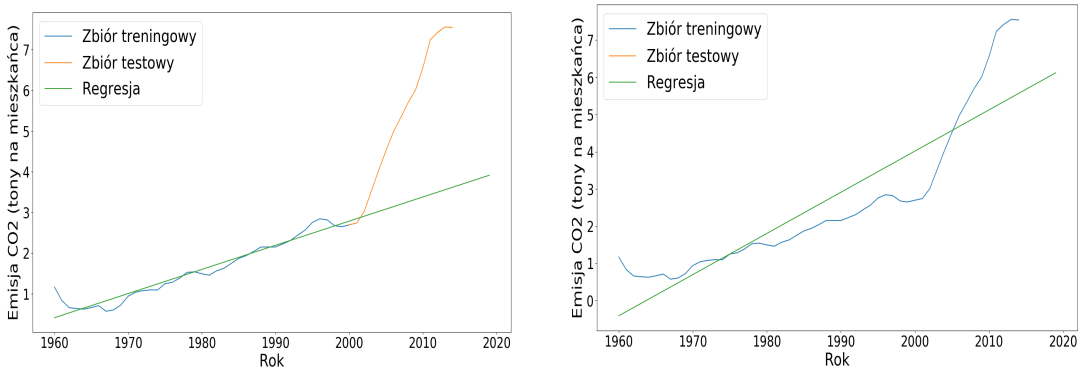



Rysunek 5: Regresja liniowa dla Maroka

Następne wykresy przedstawiają emisję dwutlenku węgla dla Chin. Jest to dość ciekawy przypadek jej gwałtownego wzrostu po roku 2000. Po lewej stronie kolorem pomarańczowym zaznaczone są wartości $Y$, które nie były uwzględniane przy tworzeniu modelu regresji liniowej. Po prawej stronie, już wszystkie zmienne objaśniane zostały wzięte pod uwagę, co znacząco wpłynęło na kąt nachylenia prostej względem poziomej osi. Po zawarciu czternastu ostatnich lat w zbiorze treningowym, otrzymujemy mniejszą różnicę między faktyczną wartością $y_{55} = 7.54$ a jej estymatorem, lecz wciąż nie jest to dobra predykcja. Co więcej, model nie opisuje prawidłowo pozostałej części wykresu.

Żadna linia prosta nie jest w stanie przedstawić charakterystyki tego wykresu, zatem warto spróbować dobrać inną regresję.

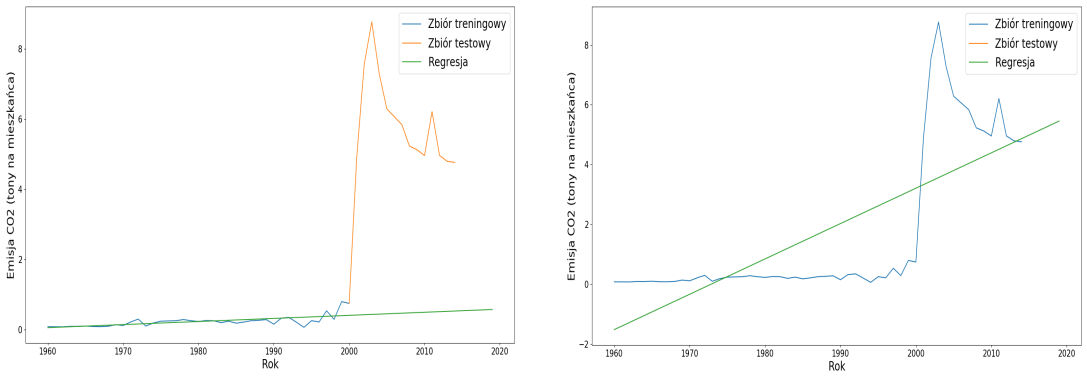
Przypomnijmy, że dla Polski $y_{55} = 7.52$. Zatem w 2014 roku emisja $CO_2$ dla Chin i Polski była niemalże taka sama, gdy jeszcze w 1980 różniła się o ponad 11.5 tony.




Rysunek 6: Regresja liniowa dla Chin

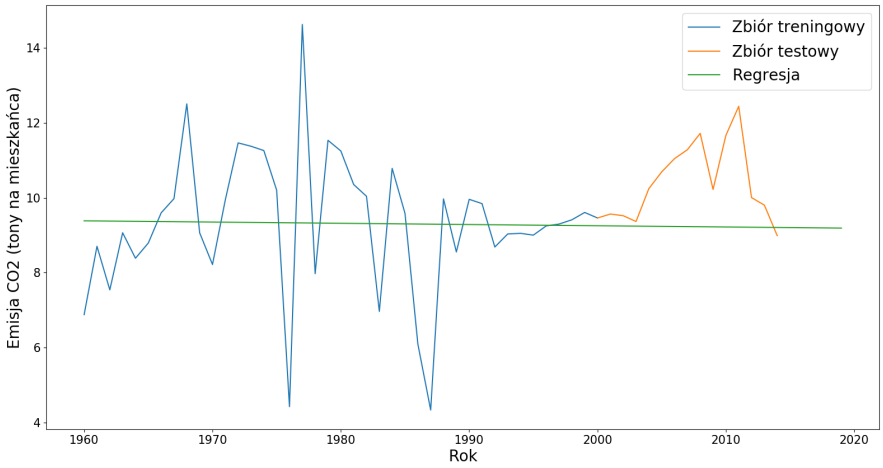
Podobny kształt wykresu emisji do Chin przybiera Gwinea Równikowa (niewielkie państwo położone w centralnej Afryce). Jednak nagły wzrost przed 2000 rokiem jest nawet bardziej widoczny, a końcowy fragment krzywej jest malejący. Początkowo państwo to opierało swoją gospodarkę na rolnictwie. W 1996, gdy odkryto złoża ropy naftowej, rozpoczął się potężny wzrost gospodarczy oraz 1170% wzrost emisji $CO_2$ w ciągu trzech lat.

Po lewej stronie linia regresji jest stale bliska zeru. Jest ona bardzo oddalona od wartości $Y$ spoza zbioru treningowego. Natomiast po włączeniu do niego pozostałych wartości $Y$, linia znacząco zmieniła kąt nachylenia względem zmiennych $X$. Ta zmiana sprawiła, że model idealnie estymuje ostatnią znaną wartość $Y$. Jednak tak jak w przypadku wykresu emisji w Chinach, Żadna linia nie będzie dobrym estymatorem ani predyktorem, zatem warto spróbować dobrać inny model.




Rysunek 7: Regresja liniowa dla Gwinei Równikowej

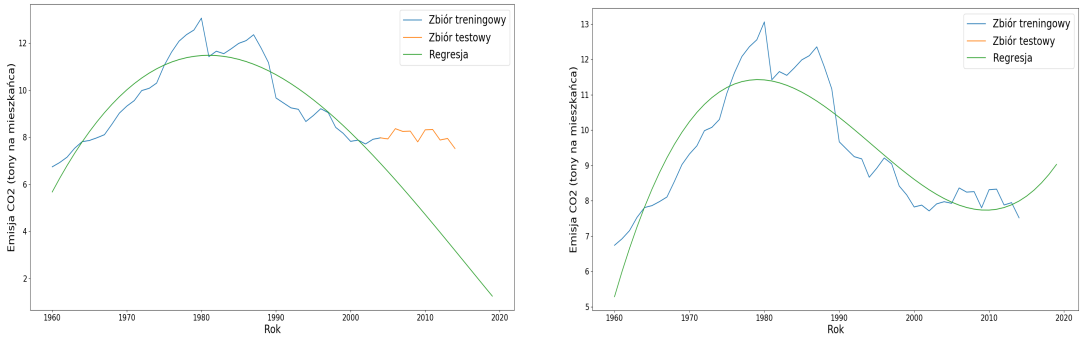
Ostatni wykres przedstawia emisję dwutlenku węgla na jednego mieszkańca Grenlandii. Jak widać prosta regresji jest prawie równoległa do osi $X$, zatem przewiduje stałą wartość $Y$ bliską liczbie 10. Tymczasem wartości tej zmiennej do roku 2000 bardzo się wahają. Z jednej strony model nie przedstawia zmienności emisji. Z drugiej, dobrze przewiduje wartość z roku 2014 oraz nie jest nadmiernie dopasowany, a to znaczy że nie wpasowuje się w każdą zmianę czy błąd lecz zachowuje zdolność generalizowania, przez co lepiej sobie radzi z danymi ze zbioru testowego. Możemy natomiast podejżewać, że model jest niedotrenowany, czyli nie przedstawia żadnych cech wykresu, powodując pewną losowość w predykcji.




Rysunek 8: Regresja liniowa dla Grenlandii

## 3.2 Model wielomianowy

Przewidujemy, że funkcje wielomianowe lepiej niż proste linie dopasują się do niemonotonicznych danych. Jednakże są bardzo wrażliwe na wartości odstające. Zobaczmy, jakie mogą być skutki stosowania tej regresji.



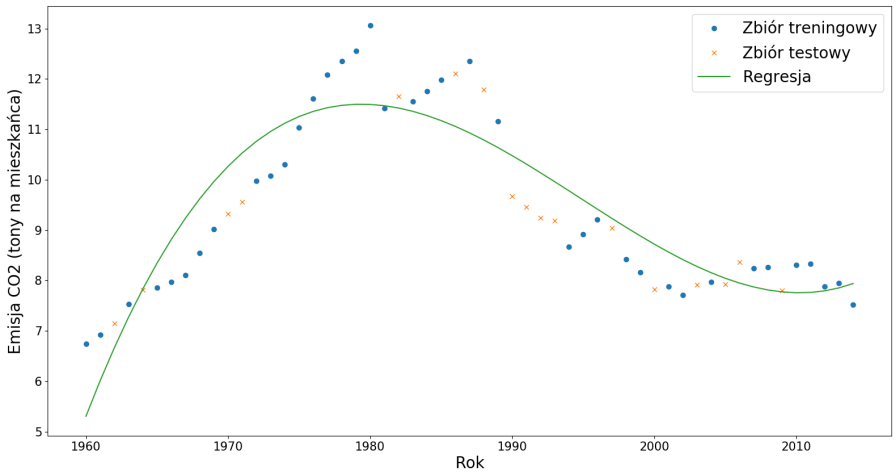

Rysunek 9: Regresja wielomianowa dla Polski

Spójrzmy na wykresy Polski z naniesioną krzywą regresji wielomianowej. Próba przewidywania wartości $Y$ w 2014 roku, przy wykorzystaniu wartości zmiennej zależnej odpowiadającej latom 1960-2000 nie daje zadowalających efektów. Algorytm użyty w predykcji, pochodzący z biblioteki *scikit-learn*, dobrał stopień wielomianu równy dwa (mimo braku jego ograniczeń). Dużą wadą takiego działania jest spotęgowany spadek wartości powodujący, że emisja przekracza dolną granicę wartości $Y$ równą zero. W przypadku testowania modelu, na wszystkich wartościach zmiennej zależnej, algorytm dobrał stopień równy 3, lepiej dopasowując się do wykresu. Widać natomiast przy krańcach przedziału, że dalesze przewidywania nie byłyby

wiarygodne.

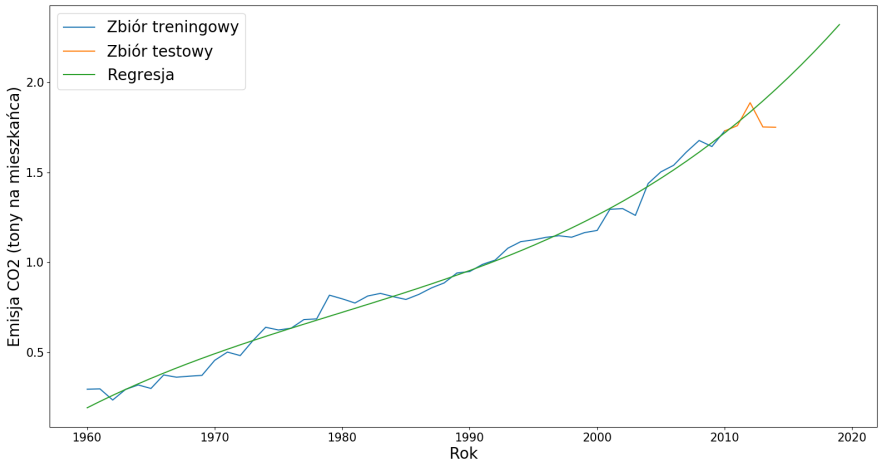
Dla porównania, model liniowy uczony na tych samych zbiorach, różnił się trendem. Dla mniejszego zbioru treningowego był to trend rosnący, dla większego zbioru - trend malejący. Niemniej jednak obie te funkcje nie odbiegały znacząco od poziomej lini.

Zbiory możemy też dzielić w inny sposób, na przykład losując 70% wartości $Y$ do zbioru treningowego, bez względu na zmienną $X$. Pozostałe wartości tworzą zbiór testowy. Tak też zbudowany jest poniższy model, który jak widać nie różni się znacząco od funkcji wielomianowej dopasowanej do pełnego modelu. Jednak krzywa na końcu przedziału wolniej rośnie, znacznie lepiej dopasowując się do danych.



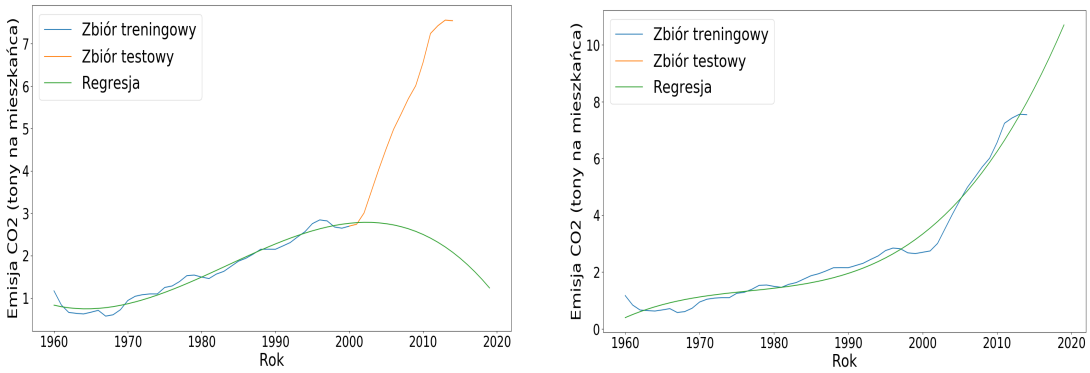

Rysunek 10: Regresja wielomianowa dla Polski dla losowego zbioru treningowego

W przypadku Maroka, funkcja wielomianowa bardzo dobrze wpasowała się w wartości zmiennej odpowiedzi, gdy zbiorem treningowym były dane z lat 1960-2010. Funkcja ta jest bardzo łagodna, przypominająca prostą linię, która również bardzo dobrze estymowała wartości zmiennej.



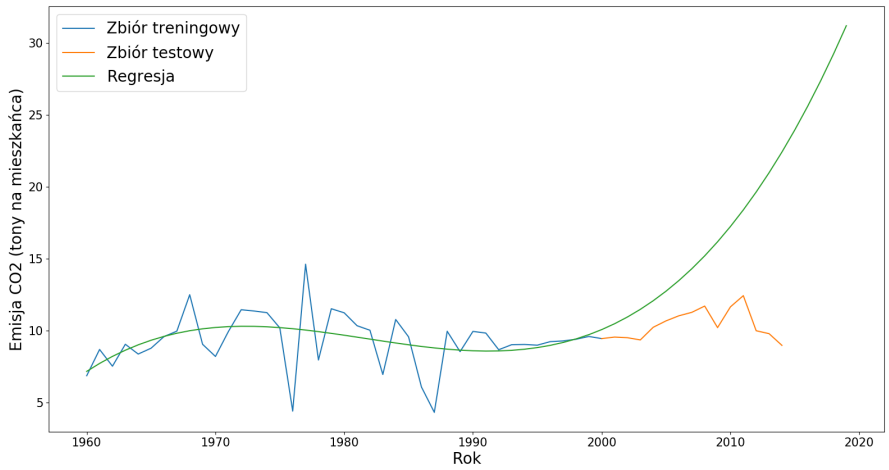

Rysunek 11: Regresja wielomianowa dla Maroka

Dla Chin podział na zbiory treningowe i testowe ma duże znaczenie. Gdy model przewiduje wartości z lat 2000-2014 (nie dostając informacji o gwałtownym wzroście emisji), kieruje się ku dołowi i estymuje $y_{55}$ na wartość bliską 1, gdy prawdziwa wartość jest bliska 7. Jest to przykład na to, że przy niewłaściwym zbiorze treningowych, wartości estymowane mogą się znacząco różnić od wartości testowych. Na drugim wykresie, cały zbiór $Y$ jest zbiorem treningowym, a model regresji bardzo dobrze reprezentuje dane. Problemem jest tu coraz szybszy wzrost krzywej. Zatem dalsza predykcja mogłaby okazać się nietrafiona.



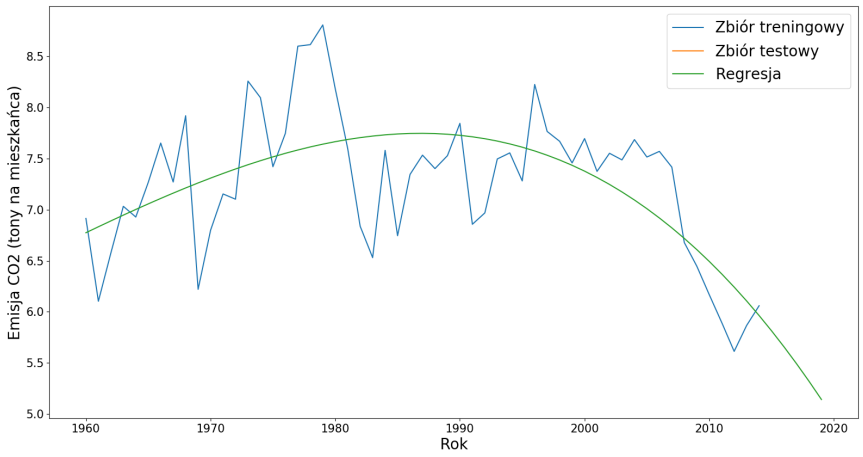

Rysunek 12: Regresja wielomianowa dla Chin

Estymator $Y$ w regresji wielomianowej bardzo szybko rośnie spłaszczając wykres emisji dwutlenku węgla na mieszkańca Grenlandii. Jego przewidywana wartość dla roku 2020 jest dwukrotnie większa od maksymalnej wartości $Y$. Do roku 2000 model bardzo dobrze estymował wartości zmiennej objaśnianej, lecz potem zaczął gwałtownie rosnąć, znacząco oddalając się od prawdziwych danych.




Rysunek 13: Regresja wielomianowa dla Grenlandii

Tym razem estymujemy model dla wszystkich wartości emisji dwutlenku węgla na Islandii. Po dokładniejszemu przyjrzeniu się zmiennym zależnym, możemy znaleźć pewną okresowość. Jednak funkcja wielomianowa uogólnia dane, estymując je jednym zakrzywieniem. Pokrywa się z początkową i końcową wartością $Y$. Po roku 2000 funkcja znacząco maleje.




Rysunek 14: Regresja wielomianowa dla Islandii

### 3.3 Wybór modelu

Regresja wielomianowa, jak podejrzewaliśmy na początku tego rozdziału, lepiej oddaje charakter zmiennej objaśnianej, jednak niepokojące jest jak zazwyczaj dąży ku górze lub ku dołowi na końcach przedziału. Jednak narzędziem, które posłuży nam do porównania modeli będzie $MSE$. Jest to miara, która ocenia dopasowanie funkji, a nie jej kierunek dążenie poza przedziałem.

Zacznijmy od regresji wykonanej dla Polski. Poniżej znajduje się tabelka z $MSE$ dla regresji liniowej i wielomianowej oraz dla zbiorów treningowych. Lata w nagłówkach kolumn odpowiadają ostatnim wartościom $Y$ włączonym do zbioru treningowego.

| Regresja/Rok | 2004 | 2014 |
|---|---|---|
| Liniowa | 3.23 | 2.88 |
| Wielomianowa | 2.47 | 0.55 |

Model liniowy trenowany na niepełnym zbiorze $Y$ ma $MSE$ równe 3.23. Jest to największa wartość tabelki zatem jest to także najgorsza predykcja według błędu średniokwadratowego. Modelem, który wypadł najlepiej jest model wielomianowy na pełnym zbiorze $Y$. Analiza wykresu potwierdza ten wniosek, ponieważ model dobrze oddaje charakter danych, poza pierwszymi kilkoma wartościami $\hat{Y}$.

W obu przypadkach regresji, $MSE$ na pełnym zbiorze wypadało lepiej. Taka zależność będzie prawdziwa w każdym kolejnym porównaniu.

Wartości $MSE$ dla regresji emisji $CO_2$ w Maroko trenowane na latach 1960-2010 oraz 1960-2014 są bardzo małe. Jest to spowodowane prawie liniowym rozkładem wartości $Y$. Wszystkie te wartości były nie większe niż 0.007. Jest to bardzo dobry wynik dla badanej miary.

Kolejna tabelka przedstawia wartości błędu dla Chin.

| Regresja/Rok | 2000 | 2014 |
|---|---|---|
| Liniowa | 1.76 | 0.76 |
| Wielomianowa | 2.95 | 0.09 |

Wraz ze wzrostem wielkości zbioru treningowego, $MSE$ zmalało o ponad 55% w modelu liniowym oraz o 96% w modelu wielomianowym. Najmniej dopasowanym modelem jest wielomianowy, uczony na niepełnym zbiorze zmiennej $Y$, natomiast najlepiej-model wielomianowy uczony na zbiorze 55 obserwacji.

Testowane na zbiorach tej samej wielkości były również modele dla Grenlandii oraz Islandii. I tak samo jak dla Chin, najmniej dopasowanym modelem był wielomianowy trenowany na zbiorze odpowiadającym latom 1960-2000. Ta wartość kilkukrotnie przewyższa wszystkie pozostałe i pokazuje wadę regresji wielomianowej, którą są gwałtowne spadki lub wzrosty przy krańcach. Według wartości błędu, regresja wielomianowa na pełnym zbiorze jest dobrze dopasowana.

Mimo, że wykresy emisji (na przykład numer 13 i 14) oraz tabelki $MSE$ obu tych państw mają podobną charakterystykę, to bardzo różnią się wielkościami błędu.

| Regresja/Rok | 2000 | 2014 |
|---|---|---|
| Liniowa | 3.31 | 3.07 |
| Wielomianowa | 11.82 | 2.87 |

| Regresja/Rok | 2000 | 2014 |
|---|---|---|
| Liniowa | 0.68 | 0.47 |
| Wielomianowa | 2.33 | 0.28 |

Porównując wartości błędu średniokwadratowego ze wszystkich przedstawionych Państw, bezkonkurencyjne okazały się wszystkie modele Maroka. Potwierdzeniem są wykresy (na przykład numer 5 i 11), do których krzywe regresji bardzo dobrze się dopasowywały.

Największą wartość $MSE$ osiągnął model wielomianowy dla Grenlandii uczony na latach 1960-2000. Tu również wykres numer 13 dobrze obrazuje wady tego modelu.

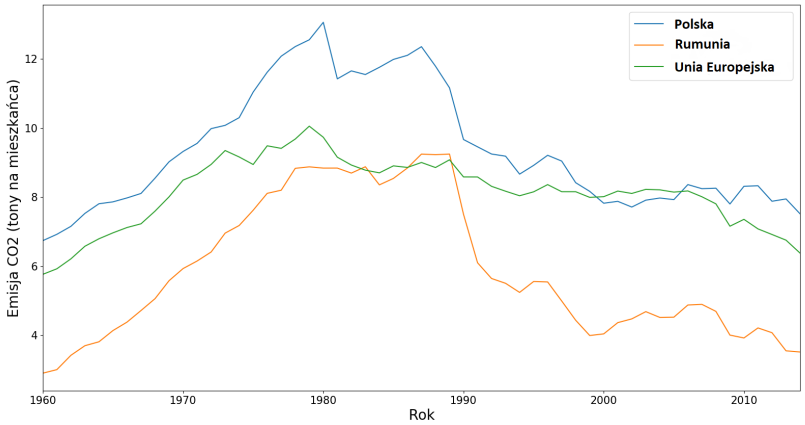
Regreja wielomianowa uczona na wszystkich wartościach $Y$ okazała się najlepsza dla każdego kraju. Na tej podstawie można powiedzieć, że jest to najlepszy wybór do estymacji emisji $CO_2$.

### 3.4 Analiza korelacji

Większość języków programowania oblicza korelację na podtawie współczynnika Pearsona, tak też działa specjalna funkcja w bibliotece $scikit - learn$. Ciekawym przykładem pokazania silnej korelacji, mimo sporego przesunięcia pionowego, jest związek Rumunii i Polski. Współczynnik korelacji Pearsona wynosi 0.97, zatem zmienne $Y_P$ (zmienna objaśniana dla polski) oraz $Y_R$ (zmienna objaśniana dla Rumunii) są mocno zależne od siebie, mimo stałej różnicy między $i$-tymi wartościami, wynoszącej około 4 tony.

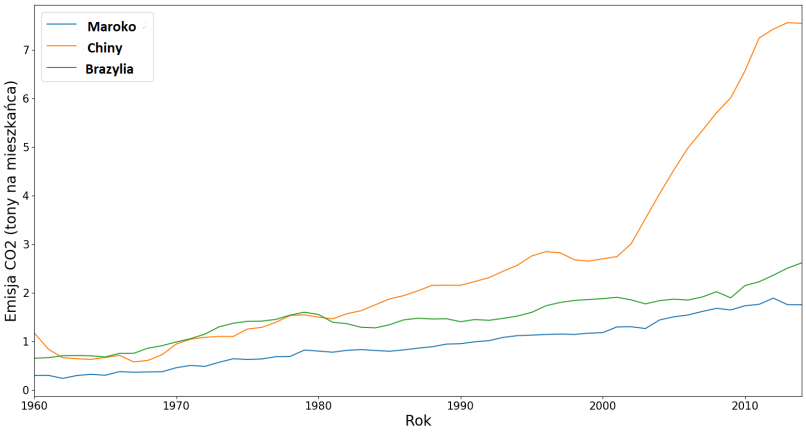
Jeżeli chodzi o stopień reprezentatywny Polski w Uni Europejskiej, to według współczynnika korelacji jest to 0.85, czyli dość sporo, lecz nie na tyle by na podstawie emisji dwutlenku węgla w Polsce móc dokładnie opisać jego emisję w Uni Europejskiej.

Przypomnijmy, że współczynnik korelacji Pearsona jest przeznaczony dla zmiennych liniowych, zatem stopień zależności wyznaczany tym sposobem może nie być wiarygodny. Niemniej jednak, stopnie skorelowania wydają się bardzo dobrze opisywać prawdziwą zależność.




Rysunek 15: Wykresy dla: Polski, Rumunii oraz Uni Europejskiej

Kolejny zestaw zawiera wykresy Brazylii ,Maroka i Chin. Dwa pierwsze mają współczynnik wzajemnej korelacji równy 0.95, co faktycznie jest zauważalne. Co ciekawe współczynnik dla Chin i Maroka wynosi 0.94, a wykres pierwszego z nich, w pewnym momencie znacznie oddala się od drugiego. Natomiast współczynnik dla Chin i Brazylii wynosi już zaledwie 0.88. Na przykładzie wykresu, obie zależności wydają się być mniejsze. Jeżeli chodzi o zależność poszególnych wykresów od lat to jest to dla Maroka, Brazylii i Chin kolejno 0.99, 0.94, 0.89. Faktycznie wszystkie te wykresy w miarę monotonicznie rosną wraz ze wrostem zmiennej objaśniającej. Dla porównania wykresy Polski i Rumunii miały współczynnik korelacji Pearsona w stosunku do lat równy -0.2 zatem mają raczej przeciwne zależności. Z kolei dla wykresu Uni Europejskiej jest to -0.017, czyli jest to dość neutralna zależność, ponieważ wartość bezwzględna tej liczby jest bliska zeru.




Rysunek 16: Wykresy dla: Maroka, Chin oraz Brazylii

# 4 Biblioteka *scikit-learn*

W języku Python można znaleźć wiele bibliotek służących do analizy statystycznej. Jedną z nich jest *scikit-learn*, zbudowana na *NumPy, SciPy*, czy *matplotlib*. Posiada ona wydajne, proste i numerycznie poprawne narzędzia do predykcji danych, klasyfikacji, grupowania, redukcji wymiarów czy wyboru modelu regresji. Jest zatem doskonałym narzędziem do analizy danych i ich obróbki.

Modułem odpowiedzialnym za stworzenie modelu liniowego jest *sklearn.linear_model.LinearRegression*. Zawiera on metody takie jak *fit* - dobierającą model do danych, czy *predict* - przewidującą kolejne wartości zmiennej objaśnianej. Pomocne są także metody obliczające $MSE$ lub $R^2$.

## Literatura


[1] Weihs, Claus, Ickstadt, Katja *Data Science: the impact of statistics. International Journal of Data Science and Analytics* (2018).

[2] https://data.worldbank.org/indicator/EN.ATM.CO2E.PC

[3] https://pl.wikipedia.org/wiki/Tendencja_rozwojowa

[4] https://pl.wikipedia.org/wiki/Współczynnik_korelacji_Pearsona

[5] https://pl.wikipedia.org/wiki/Zależność_zmiennych_losowych

[6] https://www.statsoft.pl/textbook/stathome_stat.html

[7] https://pl.wikipedia.org/wiki/Regresja_logistyczna

[8] https://towardsdatascience.com/data-science-explaining

[9] Statistics LibreTexts *Types of Outliers in Linear Regression* (Jun 5, 2019)

[10] Steel, Robert G. D., Torrie, James H., McGraw-Hill. *Principles and Procedures of Statistics, with Special Reference to Biological Sciences.*(1960)

[11] Bewick V, Cheek L, Ball J. *Statistics review 7: Correlation and regression.* Crit Care (2003 Dec;7(6):451-9. Epub 2003 Nov 5.)

[12] Agnieszka Nowak-Brzezińska *Regresja liniowa oraz regresja wielokrotna w zastosowaniu zadania predykcji danych* Wykład III-VI

[13] http://www.math.uni.wroc.pl/∼mbogdan/Modele_Liniowe/Wyklady/

[14] http://zsi.tech.us.edu.pl/∼nowak/odzw/korelacje.pdf

[15] https://en.wikipedia.org/wiki/Linear_regression

[16] https://pl.wikipedia.org/wiki/Kwartet_Anscombe