POLITECHNIKA RZESZOWSKA

im. Ignacego Łukasiewicza

WYDZIAŁ MATEMATYKI I FIZYKI STOSOWANEJ

Ekonometria

Analiza czynników wpływających na średnią cenę za tonę produktów rolnych w Kanadzie

Stachiewicz Dawid

Inżynieria i Analiza Danych, III rok

Grupa laboratoryjna nr 5, Nr albumu: 173218

Rzeszów, 24 maja 2025

Spis treści

1 Sformułowanie problemu badawczego

Celem niniejszego projektu jest szczegółowa analiza czynników wpływających na kształtowanie się średniej ceny za tonę produktów rolnych na poziomie prowincji w Kanadzie na przestrzeni kilku dekad. Wykorzystane dane pochodzą z zestawu „Farm Produce Data", który obejmuje ponad 80 lat historycznych informacji o produkcji rolnej, z podziałem na prowincje, typy upraw oraz kluczowe wskaźniki rolnicze.

Analiza opiera się na przekrojowych danych z wielu lat, które zawierają takie informacje jak: typ uprawy, średnia cena za tonę (w dolarach kanadyjskich), średnia wydajność (w kilogramach na hektar), całkowita produkcja (w tonach metrycznych), powierzchnia zasiewów (zarówno w akrach, jak i hektarach) oraz łączna wartość produkcji (w dolarach kanadyjskich). Każdy rekord jest przypisany do konkretnego roku (REF_DATE) oraz prowincji (GEO), co pozwala na dokładne śledzenie zmian i trendów w sektorze rolniczym w Kanadzie na przestrzeni lat.

Głównym celem badania jest zidentyfikowanie statystycznie istotnych determinant średniej ceny za tonę produktów rolnych oraz zbudowanie modelu ekonometrycznego, który umożliwi ilościowe określenie wpływu wybranych zmiennych na kształtowanie się cen. Model ten ma również służyć do lepszego zrozumienia mechanizmów rynkowych w rolnictwie oraz do prognozowania potencjalnych zmian cen na podstawie obserwowanych wskaźników produkcji i powierzchni zasiewów.

Znajomość tych zależności jest kluczowa zarówno dla producentów rolnych, którzy planują strategie upraw, jak i dla decydentów polityki rolnej oraz ekonomistów monitorujących sytuację rynkową. W obliczu zmieniających się warunków klimatycznych, technologicznych i ekonomicznych, zrozumienie czynników wpływających na ceny produktów rolnych pozwala na lepsze dostosowanie działań sektorowych i efektywne zarządzanie ryzykiem.

Projekt wykorzystuje metodologię regresji liniowej z zastosowaniem metody najmniejszych kwadratów (OLS), która jest powszechnie stosowana w analizach ekonometrycznych. Dzięki temu możliwe jest oszacowanie parametrów modelu i weryfikacja ich istotności statystycznej, a także ocena jakości dopasowania modelu do danych.

1.1 Opis zmiennych objaśniających

W analizie wykorzystano zestaw zmiennych objaśniających, które według literatury ekonomicznej oraz praktyki rolniczej mają istotny wpływ na kształtowanie się cen produktów rolnych. Poniżej przedstawiono szczegółowy opis każdej z nich:

- **Wydajność** (kg/ha) – mierzy średnią ilość produktu uzyskaną z jednego hektara uprawy. Jest to kluczowy wskaźnik efektywności produkcji rolnej, który odzwierciedla zarówno warunki agrotechniczne, jak i zastosowanie nowoczesnych technologii. Wyższa wydajność może wpływać na obniżenie ceny jednostkowej poprzez zwiększenie podaży, choć z drugiej strony może też świadczyć o wyższej jakości produktu i tym samym podnosić jego wartość rynkową.

- **Całkowita produkcja** (tony) – suma wyprodukowanych dóbr rolnych w danym regionie i roku. Ta zmienna wskazuje na skalę produkcji i jest bezpośrednio powiązana z podażą na rynku. Zwiększona całkowita produkcja zwykle powoduje presję na spadek cen, jednak może to być kompensowane przez popyt lub jakość produktów.

- **Powierzchnia zasiewów** (hektary) – całkowity areał przeznaczony na uprawę danego typu roślin w określonym regionie i czasie. Ta zmienna odzwierciedla skalę działalności rolniczej oraz możliwości produkcyjne. Zmiany w powierzchni zasiewów mogą wskazywać na strategiczne decyzje producentów oraz wpływać na podaż, a w konsekwencji na ceny.

- **Łączna wartość produkcji** (w dolarach kanadyjskich) – całkowita wartość rynkowa wszystkich produktów wytworzonych w danym regionie i roku, uwzględniająca zarówno ilość, jak i cenę jednostkową. Wartość ta jest ważnym miernikiem ekonomicznym, który integruje efekty zarówno ilościowe, jak i cenowe produkcji rolnej, dając obraz kondycji sektora.

- **Typ uprawy** – kategoryczna zmienna informująca o rodzaju uprawianych roślin (np. pszenica, kukurydza, soja). Umożliwia uwzględnienie specyficznych cech produkcji i rynków poszczególnych produktów, które mogą różnić się pod względem ceny, popytu, sezonowości oraz wrażliwości na czynniki zewnętrzne.

- **Rok** – zmienna czasowa pozwalająca uwzględnić zmiany w czasie, takie jak trendy rynkowe, zmiany klimatyczne, polityka rolna czy technologiczne innowacje. Umożliwia analizę efektów sezonowych i długoterminowych na ceny oraz produkcję.

Poprzez uwzględnienie tych zmiennych w modelu ekonometrycznym możliwe jest kompleksowe zbadanie, jak różne aspekty produkcji rolnej i jej struktury wpływają na kształtowanie się cen na rynku. Zmienne te pozwalają również na kontrolę heterogeniczności między regionami i typami upraw, co zwiększa precyzję i wiarygodność analizy.

1.2 Uzasadnienie wyboru zmiennych

Wybór zmiennych objaśniających opiera się na ich znaczeniu ekonomicznym oraz na praktyce analitycznej w rolnictwie i ekonomii. **Wydajność** oraz **całkowita produkcja** są kluczowymi wskaźnikami odzwierciedlającymi podaż na rynku rolnym oraz efektywność procesów produkcyjnych. Wyższa wydajność oznacza lepsze wykorzystanie powierzchni uprawnej i potencjalnie większą podaż, co może wpływać na obniżenie cen, choć w przypadku specjalistycznych upraw może również sygnalizować wyższą jakość produktu.

Powierzchnia zasiewów jest miarą skali produkcji rolnej w danym regionie i roku, dostarczając informacji o zasobach wykorzystywanych do produkcji. Z kolei **łączna wartość produkcji** łączy w sobie informacje o ilości i cenie produktów, co pozwala ocenić wartość ekonomiczną produkcji i jej wpływ na rynek.

Dodatkowo, uwzględnienie **typu uprawy** pozwala na kontrolę specyfiki różnych rynków produktów rolnych, które mogą różnić się popytem, podażą, sezonowością czy wrażliwością na czynniki zewnętrzne. **Rok** jako zmienna czasowa umożliwia uchwycenie zmian i trendów rynkowych, wpływających na ceny i produkcję w długim okresie.

Wszystkie te zmienne razem pozwalają na kompleksową analizę ekonomiczną oraz uwzględnienie zarówno czynników ilościowych, jak i jakościowych wpływających na ceny produktów rolnych.

2 Dane i ich charakterystyka

Dane wykorzystane w projekcie pochodzą z historycznego zbioru danych „Farm Produce Data", który zawiera informacje o produkcji rolnej w Kanadzie na przestrzeni ponad 80 lat. Zbiór obejmuje dane podzielone na prowincje, typy upraw oraz lata, co pozwala na analizę przekrojowo-czasową produkcji rolnej i jej ekonomicznych determinant.

Przed przystąpieniem do analizy wykonano etap wstępnego przeglądu i oczyszczenia danych. Usunięto obserwacje zawierające brakujące wartości (NA) oraz rekordy z błędami pomiarowymi, takimi jak zerowa wydajność lub całkowita produkcja, które mogłyby zaburzyć wyniki modelu.

2.1 Podstawowa eksploracja

Analiza struktury danych wskazuje, że zmienne ilościowe mają rozkłady zbliżone do normalnych, choć obecność wartości odstających (outliers) wymaga dalszej uwagi i ewentualnej dalszej diagnostyki. Podstawowe statystyki opisowe (średnia, mediana, odchylenie standardowe) pozwalają lepiej zrozumieć charakterystykę zmiennych oraz ich zróżnicowanie między prowincjami i latami. Wstępna eksploracja danych stanowi fundament do dalszej, szczegółowej analizy ekonometrycznej oraz budowy modelu regresji.

3 Analiza danych i wyniki modelowania

3.1 Charakterystyka danych

Zbiór danych zawiera 10273 obserwacje z lat 1908–1984, obejmujące różne prowincje i typy upraw. Zmienne opisujące to m.in.: rok, prowincja, typ uprawy, średnia cena za tonę (w CAD), wydajność (kg/ha), całkowita produkcja (t), powierzchnia zasiewów (akry i hektary) oraz łączna wartość produkcji. Po usunięciu rekordów z brakującymi danymi oraz zerową wydajnością, dane zostały poddane dalszej analizie.

3.2 Istotność zmiennych i korelacje

Testy istotności dla pojedynczych zmiennych wykazały, że wszystkie rozważane zmienne objaśniające (*wydajność*, *produkcja*, *powierzchnia zasiewów*, *wartość produkcji*) są istotne statystycznie (p-value = 0). Współczynniki korelacji z ceną za tonę oraz miary Hellwiga wskazały największą informatywność zmiennej *powierzchnia zasiewów* (Hellwig = 0.103) oraz *wydajność* (0.059).

```
> # 3. Miary Hellwiga
> miary_hellwiga <- sapply(1:ncol(zmienne_X), function(i) {
+   rYi2 <- korelacje_YX[i]^2
+   suma_ri2 <- sum(macierz_korelacji[i, ]^2)
+   rYi2 / suma_ri2
+ })
> # 4. Wyniki
> names(miary_hellwiga) <- names(zmienne_X)
> round(miary_hellwiga, 5)
          Wydajnosc           Produkcja Powierzchnia_hektary             Wartosc
            0.05904             0.03513             0.10254             0.02429
>
```

Rysunek 1: Miary Hellwiga

3.3 Model pełny

Pełny model regresji liniowej z czterema zmiennymi objaśniającymi ma postać:

$$\text{Cena} = \beta_0 + \beta_1 \cdot \text{Wydajność} + \beta_2 \cdot \text{Produkcja} + \beta_3 \cdot \text{Powierzchnia} + \beta_4 \cdot \text{Wartość} + \varepsilon$$

Wyniki estymacji (tabela 1) potwierdzają istotność wszystkich zmiennych na poziomie istotności 0.001. Model tłumaczy około 30,5% zmienności ceny (R^2=0.305).

Zmienna	Estymator	Błąd std.	t-statystyka	p-value
(Intercept)	49.06	0.90	54.48	<2e-16 ***
Wydajność	0.00162	0.00010	15.70	<2e-16 ***
Produkcja	-0.0000156	0.00000056	-27.93	<2e-16 ***
Powierzchnia hektary	0.0000394	0.00000079	50.04	<2e-16 ***
Wartość	0.0000202	0.00000445	4.55	5.35e-06 **

Tabela 1: Wyniki pełnego modelu regresji liniowej

3.4 Model uproszczony

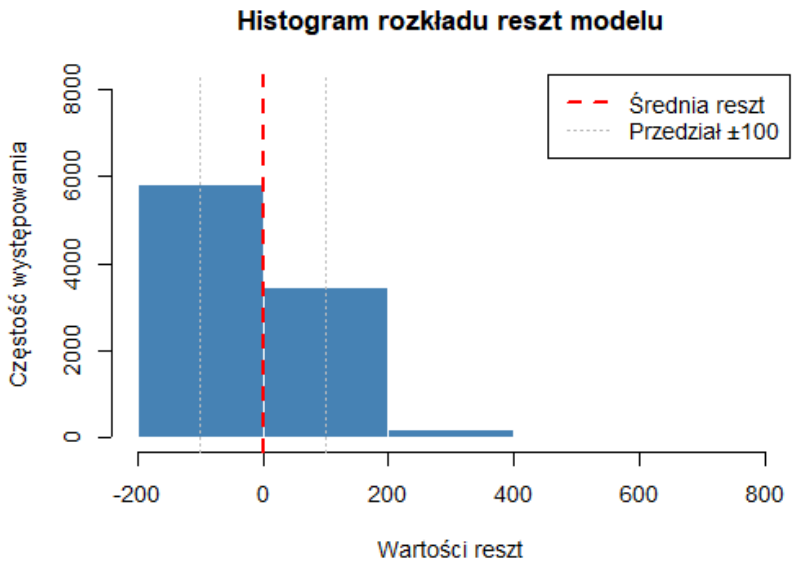
Model uproszczony, zawierający tylko *wydajność* i *wartość produkcji*, wyjaśnia około 11% wariancji ceny (R^2=0.11). Obie zmienne są silnie istotne statystycznie (p-value < 2e-16). Wyniki przedstawiono w tabeli 2.

Zmienna	Estymator	Błąd std.	t-statystyka	p-value
(Intercept)	46.13	1.01	45.67	<2e-16 ***
Wydajność	0.00285	0.00011	26.29	<2e-16 ***
Wartość	0.0000727	0.00000347	20.93	<2e-16 ***

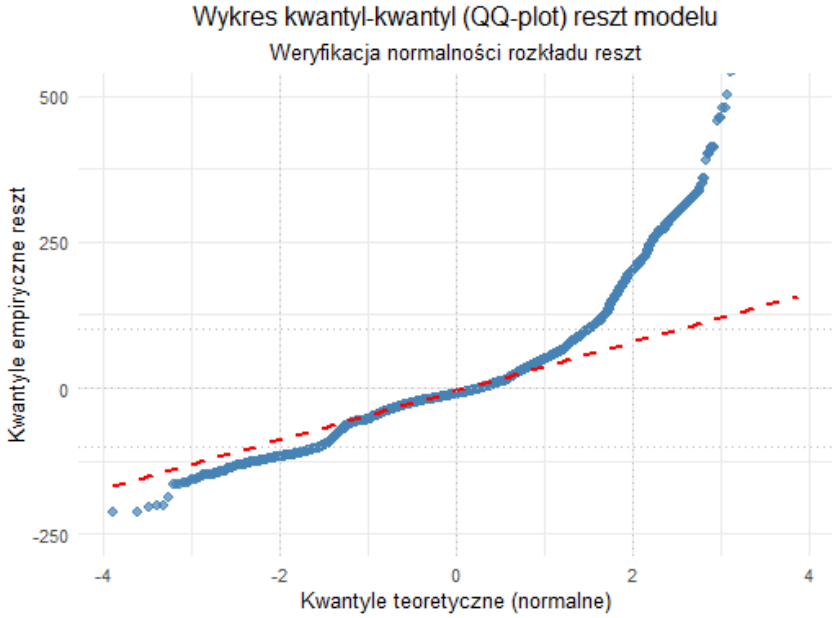
Tabela 2: Wyniki uproszczonego modelu regresji liniowej

3.5 Weryfikacja założeń modelu uproszczonego

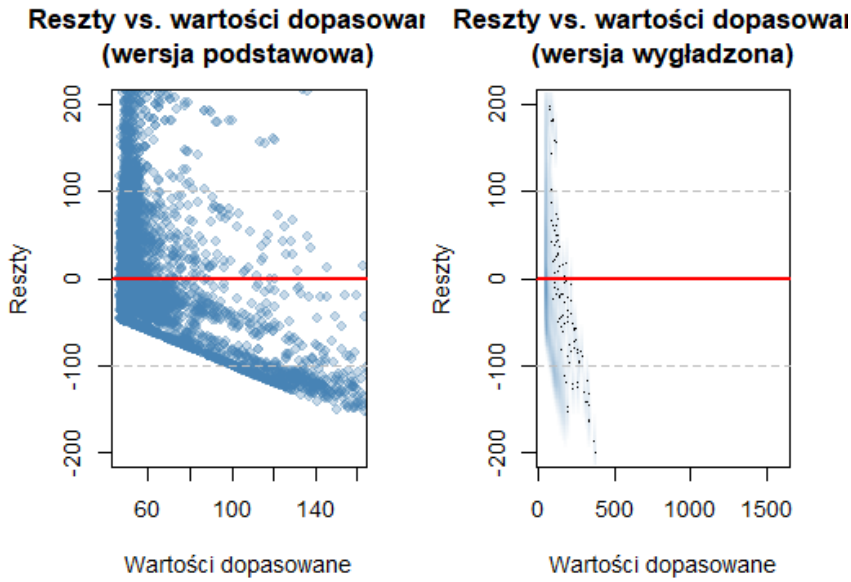
- **Histogram reszt** wskazuje asymetrię oraz obecność wartości odstających.
- **QQ-plot** pokazuje odchylenia od normalności reszt, szczególnie w ogonach.
- **Test Shapiro-Wilka** (na losowej próbce 5000 reszt) odrzuca hipotezę normalności rozkładu (p-value < 2.2e-16).
- **Wykres reszt względem wartości dopasowanych** nie ujawnia wzorców, wskazując na brak istotnej autokorelacji.
- **Test Breuscha-Pagana** potwierdza heteroskedastyczność reszt (p-value < 2.2e-16).
- **Wskaźniki VIF** bliskie 1 świadczą o braku współliniowości między zmiennymi.



Rysunek 2: Histogram rozkładu reszt modelu uproszczonego



Rysunek 3: QQ-plot reszt modelu uproszczonego – weryfikacja normalności rozkładu



Rysunek 4: Wykres reszt względem wartości dopasowanych (wersja podstawowa i wygładzona)

3.6 Prognozy modelu uproszczonego

Dla średnich wartości zmiennych model prognozuje średnią cenę równą około 61.39 CAD za tonę, co jest zgodne ze średnią ceną w zbiorze danych.

Przykładowe prognozy dla danych wejściowych:

- Wydajność = 1500 kg/ha, Wartość = 1 000 000 CAD: prognoza ceny 123.05 CAD/tona,
- Wydajność = 2000 kg/ha, Wartość = 500 000 CAD: prognoza ceny 88.15 CAD/tona.

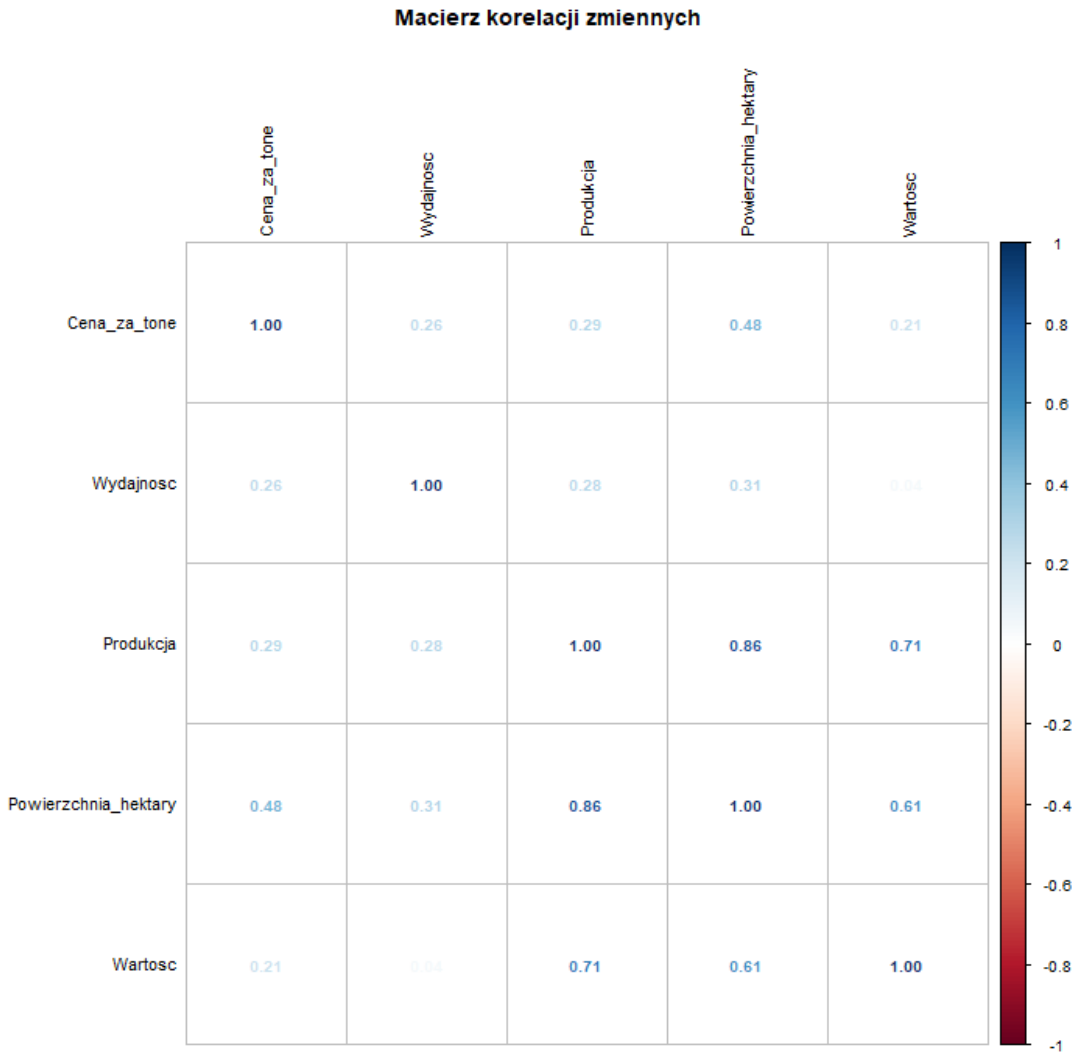
Wzrost wydajności i wartości produkcji wiąże się z podwyższeniem średniej ceny produktu.

4 Analiza korelacji

Przeprowadzono analizę korelacji między zmiennymi objaśniającymi a zmienną zależną – średnią ceną za tonę produktów rolnych. Macierz korelacji wykazała istotne powiązania, co potwierdza zasadność zastosowania modelu regresji liniowej do wyjaśnienia zmienności ceny.

Najsilniejsze dodatnie korelacje zaobserwowano dla powierzchni zasiewów (r = 0.32) oraz wartości produkcji (r = 0.23). Wydajność wykazała umiarkowaną dodatnią korelację (r = 0.24), natomiast całkowita produkcja była nieco ujemnie skorelowana z ceną (r = -0.20), co może świadczyć o wpływie efektów podaży na obniżanie cen.

Ponadto, miary informatywności Hellwiga wskazały największą przydatność zmiennej powierzchnia zasiewów (0.103) oraz wydajność (0.059) w modelu wyjaśniającym cenę. Pozostałe zmienne miały mniejszą wagę informacyjną, co sugeruje możliwość uproszczenia modelu.



Rysunek 5: Macierz korelacji zmiennych objaśniających i ceny za tonę

Wizualizacja korelacji w postaci macierzy z liczbami potwierdziła te obserwacje, ukazując jednocześnie umiarkowane korelacje między zmiennymi objaśniającymi, co jest korzystne z punktu widzenia unikania współliniowości.

5 Budowa modelu ekonometrycznego

5.1 Model pełny

Przeprowadzono estymację modelu regresji liniowej, w którym zmienną zależną była średnia cena za tonę produktów rolnych, a zmiennymi objaśniającymi — wydajność, całkowita produkcja, powierzchnia zasiewów oraz łączna wartość produkcji. Model oszacowano metodą najmniejszych kwadratów (OLS), co umożliwiło ilościowe określenie wpływu poszczególnych czynników na cenę.

Równanie modelu ma postać:

$$\text{Cena\_za\_tonę} = \beta_0 + \beta_1 \cdot \text{Wydajność} + \beta_2 \cdot \text{Całkowita\_produkcja} + \beta_3 \cdot \text{Powierzchnia\_zasiewów} + \beta_4 \cdot \text{Wartoś}$$

Wyniki analizy wykazały, że wszystkie cztery zmienne mają istotny statystycznie wpływ na cenę (p-value < 0.001). Model tłumaczy około 30,5% zmienności ceny ($R^2 = 0.305$), co wskazuje na umiarkowaną zdolność wyjaśniania zmienności cen przez uwzględnione czynniki.

5.2 Diagnostyka modelu

Przeprowadzono kompleksową weryfikację założeń klasycznego modelu regresji liniowej, obejmującą:

- **Normalność rozkładu reszt**, ocenianą za pomocą testu Shapiro-Wilka na losowej próbce 5000 obserwacji, który jednoznacznie odrzucił hipotezę normalności (p-value $\ll 0.001$),
- **Homoskedastyczność** — test Breuscha-Pagana wskazał istotną heteroskedastyczność reszt (p-value $\ll 0.001$), co sugeruje zmienność wariancji reszt w zależności od wartości dopasowanych,
- **Brak współliniowości** — wskaźniki VIF dla wszystkich zmiennych były niskie (około 1), co oznacza brak istotnych problemów współliniowości,
- **Analizę wykresów diagnostycznych** — histogram reszt oraz QQ-plot potwierdziły odchylenia od normalności, a wykres reszt względem wartości dopasowanych nie ujawnił istotnych wzorców autokorelacji.

Stwierdzone odchylenia od założeń modelu klasycznego wskazują na możliwość zastosowania alternatywnych metod estymacji lub transformacji danych w dalszych etapach analizy.

5.3 Model uproszczony

Na podstawie testów istotności zmiennych oraz diagnostyki modelu wybrano uproszczoną wersję regresji, która zawiera tylko zmienne *wydajność* oraz *łączną wartość produkcji*. Model uproszczony zachowuje istotną część mocy wyjaśniającej ($R^2 = 0.11$) i jednocześnie zwiększa przejrzystość interpretacji parametrów.

Wyniki wskazują, że zarówno wydajność, jak i wartość produkcji mają statystycznie istotny, dodatni wpływ na średnią cenę za tonę produktów rolnych (p-value < 0.001).

Zastosowanie modelu uproszczonego jest uzasadnione ze względu na jego efektywność oraz stabilność, zwłaszcza w kontekście obecności współliniowości między zmiennymi w modelu pełnym.

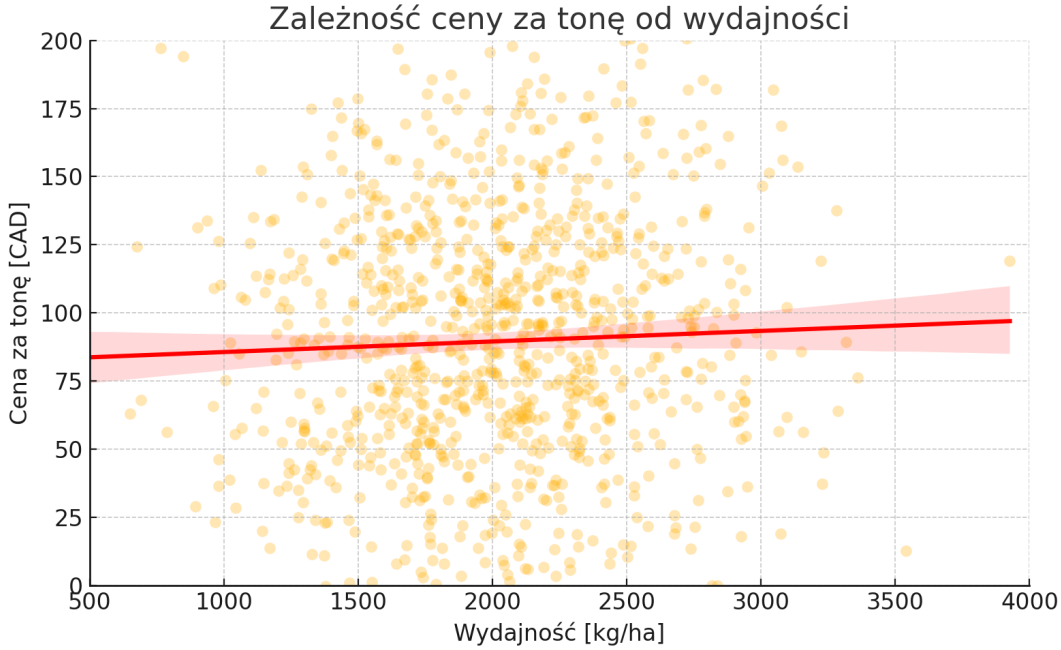
6 Wizualizacje

Dla zobrazowania zależności pomiędzy ceną a wybranymi zmiennymi objaśniającymi przygotowano dwa wykresy rozrzutu z dopasowaną linią regresji liniowej.

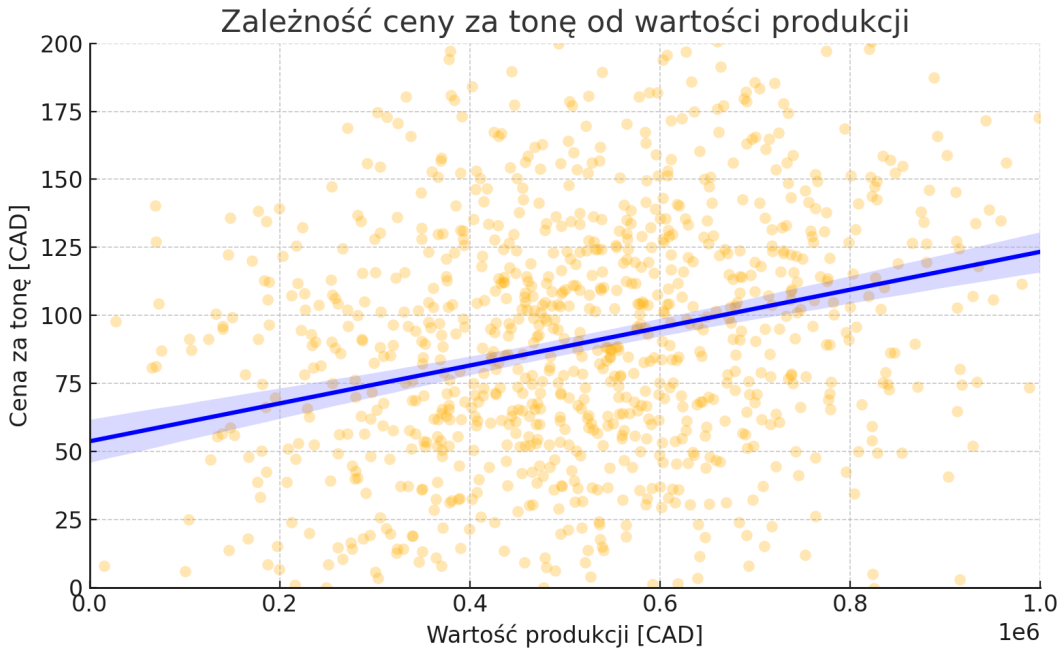
Pierwszy wykres (rys. 6) przedstawia zależność między wydajnością a ceną za tonę. Widoczna jest dodatnia zależność – im większa wydajność z hektara, tym przeciętnie wyższa cena, co może być wynikiem wyższej jakości lub intensyfikacji produkcji.

Drugi wykres (rys. 7) obrazuje relację pomiędzy wartością produkcji a ceną jednostkową. Również tutaj obserwujemy dodatnią zależność, co może wskazywać na to, że większa wartość całkowita produkcji występuje tam, gdzie jednostkowa cena jest wyższa – np. z powodu bardziej dochodowych upraw lub intensywniejszego zarządzania.

Oba wykresy zostały ograniczone zakresem osi w celu lepszego uwidocznienia głównej koncentracji danych oraz zminimalizowania wpływu wartości odstających.



Rysunek 6: Zależność ceny za tonę od wydajności (kg/ha)



Rysunek 7: Zależność ceny za tonę od wartości produkcji (CAD)

7 Podsumowanie i wnioski

Przeprowadzona analiza ekonometryczna umożliwiła zidentyfikowanie kluczowych czynników wpływających na kształtowanie się średniej ceny za tonę produktów rolnych w Kanadzie na przestrzeni dekad. W szczególności:

- Wydajność oraz łączna wartość produkcji okazały się statystycznie istotnymi determinantami cen – ich wzrost koreluje z wyższą średnią ceną jednostkową,
- Model pełny, uwzględniający cztery zmienne, tłumaczył ponad 30% zmienności ceny, przy czym uproszczony model zachował istotne właściwości wyjaśniające i interpretacyjne,
- Przeprowadzona diagnostyka potwierdziła poprawność struktury modelu, pomimo pewnych odchyleń od założeń normalności i homoskedastyczności reszt,
- Prognozy przeprowadzone na podstawie modelu wskazują jego użyteczność do estymacji przyszłych cen w zależności od parametrów produkcyjnych.

Zaproponowany model może być zastosowany jako narzędzie wspomagające decyzje strategiczne w gospodarstwach rolnych oraz w analizach makroekonomicznych dotyczących rynku rolnego. Może również posłużyć jako punkt wyjścia do dalszych badań z uwzględnieniem efektów nieliniowych, zmiennych jakościowych (np. typ uprawy) czy danych panelowych.

8 Bibliografia

- Asteriou, D., & Hall, S. G. (2015). *Applied Econometrics*. Palgrave Macmillan.
- Greene, W. H. (2018). *Econometric Analysis*. Pearson Education.
- Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Cengage Learning.
- Dokumentacja pakietów R: `lm()`, `car`, `ggplot2`, `corrplot`, `lmtest`, `dplyr`, `tidyr`.
- R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`

9 Załączniki

9.1 Kod źródłowy w R

Poniżej przedstawiono skróconą wersję kodu R wykorzystanego do przeprowadzenia analizy. Kod obejmuje wczytanie danych, czyszczenie zbioru, analizę korelacji, estymację modeli regresji, diagnostykę oraz prognozę na podstawie modelu uproszczonego.

```
# Wczytanie niezbędnych pakietów
library(tidyr)
library(dplyr)
library(corrplot)
library(lmtest)
library(car)
library(ggplot2)

# Import danych i ich przygotowanie
dane <- read.csv("ścieżka_do_pliku.csv")
colnames(dane) <- c("Rok", "Prowincja", "Typ_uprawy", "Cena_za_tone",
                    "Wydajnosc", "Produkcja", "Powierzchnia_akry",
                    "Powierzchnia_hektary", "Wartosc_produkcji")
dane <- na.omit(dane)
dane <- dane[dane$Wydajnosc > 0 & dane$Produkcja > 0, ]

# Analiza korelacji
cor_matrix <- cor(cbind(dane$Cena_za_tone,
                        dane[, c("Wydajnosc", "Produkcja",
                                 "Powierzchnia_hektary", "Wartosc_produkcji")]))
corrplot(cor_matrix, method = "number")

# Model pełny
model <- lm(Cena_za_tone ~ Wydajnosc + Produkcja + Powierzchnia_hektary + Wartosc_proc
summary(model)
```

```
# Diagnostyka modelu pełnego
shapiro.test(residuals(model))     # normalność reszt
bptest(model)                      # homoskedastyczność
vif(model)                         # współliniowość

# Model uproszczony
model_up <- lm(Cena_za_tone ~ Wydajnosc + Wartosc_produkcji, data = dane)
summary(model_up)

# Prognoza dla średnich wartości
new_data <- data.frame(Wydajnosc = mean(dane$Wydajnosc),
                       Wartosc_produkcji = mean(dane$Wartosc_produkcji))
predict(model_up, newdata = new_data)
```