Analiza wydajności krów mleczych w gospodarstwach europejskich

Mateusz Marzec, Szymon Święs

1 Wstęp

Celem badania jest przeanalizowanie czynników wpływających na wydajność krów mlecznych w gospodarstwach europejskich. Analiza ta nie jest poparta teorią ekonomiczną dotyczącą np. funkcji produkcji, dlatego spodziewamy się wyników potencjalnie nieużytecznych z punktu widzenia ekonomisty. Tego typu podejście może być jednak bardzo ciekawe i doprowadzić do niecodziennych wniosków. Obserwacjami są typowe gospodarstwa z poszczególnych regionów Europy (107 obserwacji). Jako zbiór danych wykorzystaliśmy jeden z raportów FADN (Wyniki Standardowe 2018, FADN). Zmienną objaśnianą jest "Wydajność mleczna krów", podawana w kg/krowę w ujęciu rocznym (oznaczenie SE125). Jest to produkcja mleka i przetworów z mleka (w ekwiwalencie mleka) w przeliczeniu na jedną krowę mleczną. Produkcja mleka obejmuje sprzedaż, przekazanie mleka do gospodarstwa domowego i zużycie mleka w gospodarstwie rolnym (skarmione przez zwierzęta).

Zmienne objaśniające:

1) Dopłaty do pozostałej produkcji zwierzęcej [zł] (SE615) - W jej skład wchodzą dopłaty uzyskiwane z tytułu chowu oraz produkcji bydła. Zależy nam na zbadaniu czy różnego rodzaju dofinansowania mają wpływ na jakość produkcji mleka. Każda ze zmiennych jest podawana w ujęciu rocznym.

2) Uprawy pastewne [ha] (SE071) - Wielkość upraw pastewnych korzeniowych i kapustnych, traw w uprawie polowej, łąk, pastwisk trwałych i niepielęgnowanych.

3) Mleko i przetwory z mleka krowiego [zł] (SE216) - Jest to wartość mleka i przetworów pomniejszona o odpowiednie opłaty współodpowiedzialnościowe, lecz bez opłat karnych za nieprawidłową realizację kwoty mlecznej.

4) Pasze własne [zł/krowa] - wydatki na pasze wyprodukowane w gospodarstwie przypadających na jedną krowę, zmienna obliczna własnoręcznie.

5) Pasze obce [zł/krowa] - wydatki na pasze zakupione przez gospodarstwa przypadające na jedną krowę, zmienna obliczna własnoręcznie.

6) Koszty bezpośrednie produkcji roślinnej [zł/ha] (SE284) - Jest to wartość kosztów bezpośrednich produkcji roślinnej w przeliczeniu na 1 ha użytków rolnych. Będziemy interpretowali tą zmienną jako wartość inwestycji gospodartw w produkcję roślinną.
7) Nakłady pracy ogółem [AWU-2120 h/rok] (SE010) - Całkowite nakłady pracy ludzkiej w ramach działalności operacyjnej gospodarstwa rolnego. Podawane w jednostak AWU, umownej jednostce nakładów pracy w rolnictwie. Klasyczna zmienna ekonomiczna.
8) Pozostałe koszty bezpośrednie produkcji zwierzęcej (zł) [SE330] - Opłaty za usługi weterynaryjne i koszty inseminacji, koszty analiz mleka, okazjonalne zakupy produktów pochodzenia zwierzęcego, koszty dotyczące przygotowania produktów do sprzedaży, koszty przechowywania, koszty sprzedaży produktów zwierzęcych. Zmienną tą będziemy interpretować jako koszty związane z produkcją zwierzęcą.
9) Gospodarstwo (na podstawie SE005) - Jest to zmienna kategorialna reprezentująca wielkość ekonomiczną gospodarstwa, zgodna z podziałem zaproponowanym przez FADN.

2 Analiza danych

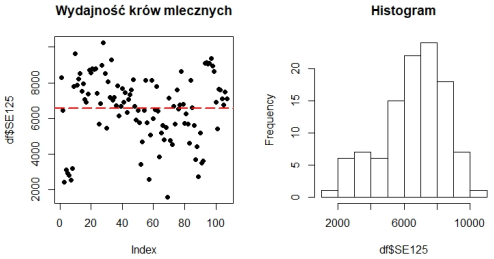
2.1 Wartości brakujące

W oryginalnym zbiorze danych występują braki danych, w postaci pustcych obserwacji (brak danych o regionie). W związku z tym usunęliśmy wiersze nic nie wnoszące do naszej analizy. Inne braki danych nie wystąpiły.

2.2 Wizualizacja

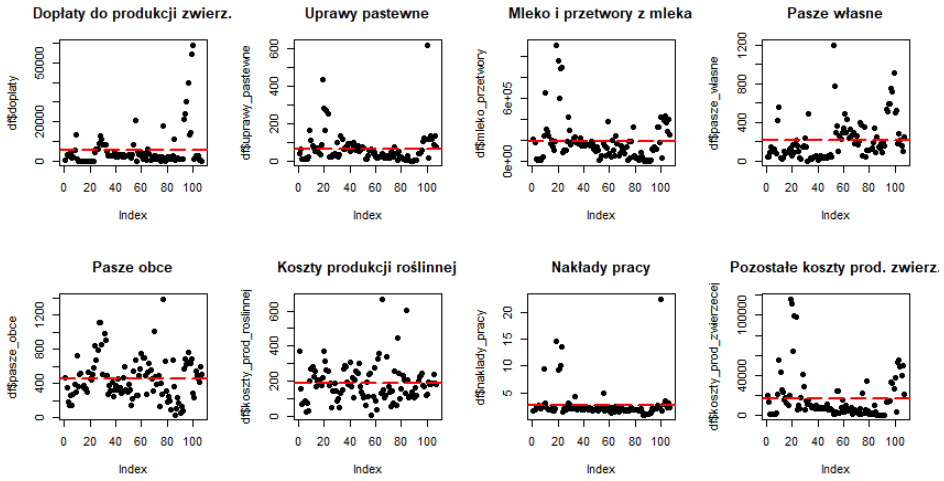
Zaczniemy od wizualnego przedstawienia poszczególnych zmiennych i wniosków, które możemy na jej podstawie wyciągnąć. (średnia oznaczana będzie przez czerwoną przerywaną linię).

Wydajność krów dla poszczególnych regionów (na podstawie rysunku 1) wydaje się zróżnicowana. Rozkład jest lewostronnie asymetryczny, większość obserwacji przyjmuje wartości większe od średniej (która w tym przypadku wynosi 6584 kilogramów mleka na jedną krowę).



Rysunek 1: Wydajność krów mlecznych (SE125)

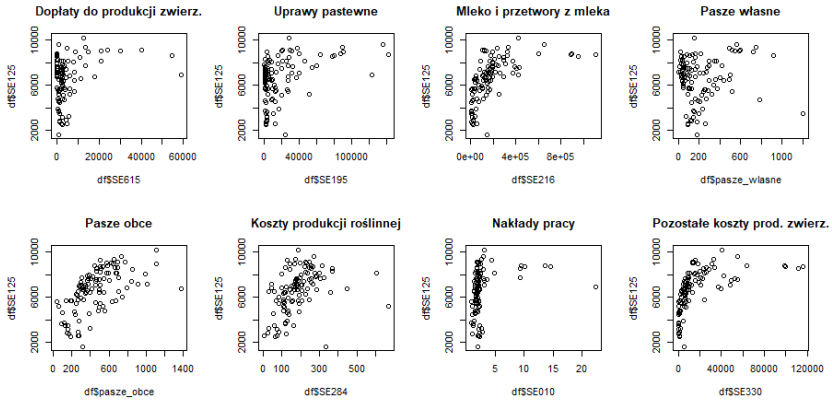
Poniżej prezentujemy zmienne objaśniające.



Rysunek 2: Wizualizacja poszczególnych zmiennych objaśniających

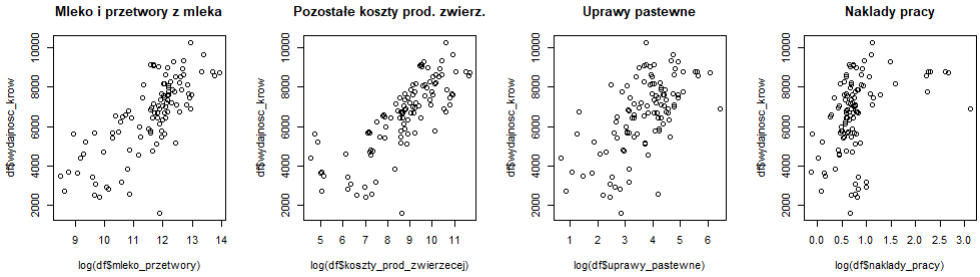
Praktycznie każda zmienna z rysunku 2 posiada obserwacje odstające. Warty podkreślenia jest fakt, że nierzadko pojawiają się one dla tych samych oberwacji. Prawdopodobnie jest to spowodowane występowaniem dużych gospodarstw, produkujących i wytwarzających spore ilości dóbr. Część uwzględnionych przez nas zmiennych ma bardzo podobny rozkład. Przykładem na to są wykresy "Mleko i przetwory z mleka" oraz "Pozostałe koszty produkcji zwierzęcej". Warto zaznaczyć też, że w przeciwieństwie do zmiennej objaśnianej rozkłady te są prawostronnie skośne (na podstawie rysunku 2). W późniejszej części pracy zbadamy czy zmienne te nie niosą tej samej informacji.

Poniżej przedstawiamy wykresy zależności pomiędzy zmiennymi.



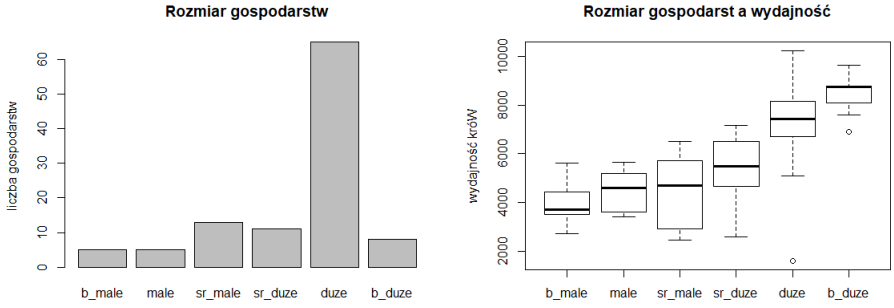
Rysunek 3: Wpływ poszczególnych zmiennych objaśnianych na wydajność krów

Z rysunku 3 wynika, że duża liczba analizowanych przez nas zmiennych w bardzo podobny sposób wpływa na zmienną objaśnianą. Część zależności (np. dla "Pozostałych kosztów produkcji" oraz "Mleka i przetworów z mleka") bardziej przypomina zależność logarytmiczną niż liniową. Jest to sugestia, aby później w modelu wykorzystać logarytm tychże zmiennych. Również z punktu widzenia ekonomicznego, logarytmy są mile widziane, ponieważ pozwalają tworzyć bardziej użyteczne, z ekonomiczego punktu widzenia interpretacje. Jedna ze zmiennych - "dopłaty" posiada zerowe wartości stąd nie będziemy zamieniać jej na logarytm.



Rysunek 4: Logarytmy zmiennych, a wydajność krów

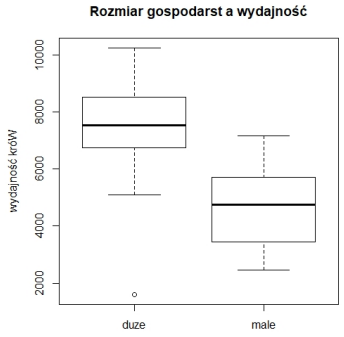
Rysunek 4 pokazuje, że transformacja zmieniła zależność na "bardziej" liniową. Na etapie budowy modelu oprócz podstawowych zmiennych uwzględnimy też ich logarytmy.
Chwilę uwagi warto również poświęcić jedynej zmiennej kategorialniej występującej w naszym zestawie danych. Jej rozkład i związek z wydajnością krów pokazuje rysunek 5.



Rysunek 5: Gospodarswa zgrupowane według wielkości ekonomicznych (według podziału zaproponowanego przez FAND)

Zecydowana większość gospodarst została zakwalifikowana jako gospodarstwa "duże". Rodzi to wątpliwości, czy ten podział ma w ogóle sens. Gdy spojrzymy na wydajność krów w obrębie poszczególnych kategorii, widać różnice pojawiające się pomiędzy poszczególnymi grupami. Wynika z nich, że wraz ze wzrostem rozmiaru gospodarstwa rośnie także wydajność hodowanych na nim krów. Budzi to nadzieję, że zmienna będzie niosła ze sobą wartość prognostyczną.

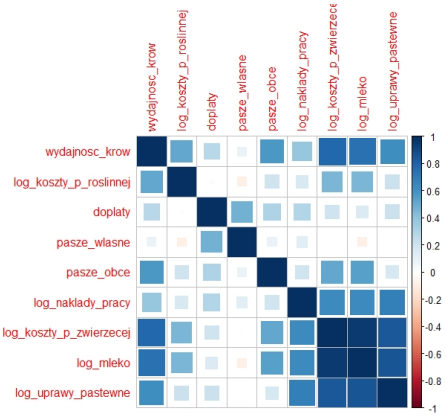
Obawy może budzić fakt, że większość gospodarstw należy do klasy "duże", przez co pozostałe klasy są mało reprezentatywne. Z tego powodu zredukujemy liczbę klas do dwóch ("małe" i "duże"), patrz rysunek 6. Zmienna ta od teraz będzie nazwana gospodarstwo_2. Patrząc na obserwacje odstające pojawiające się dla poszczególnych zmiennych na rysunku numer 2 możemy wysnuć wniosek o gospodarstawch dominujących w niektórych aspektach (jak nakłady pracy czy wielkości upraw). Wprowadzimy zatem do modelu kombinacje zmiennej informującej nas o wielkości ekonomicznej gospodarstwa i wybranych zmiennych występujących na rysunku numer 2.



Rysunek 6: Zmodyfikowana zmienna gospodarstwo

2.3 Korelacja danych wejściowych

Brak współliniowości zmiennych jest jednym z założeń modelu regresji liniowej i jego spełnienie testować będziemy w rodziale 4. Przed wyborem zmiennych objaśniających czasem warto przeanalizować macierz korelacji. Często pomaga to w wyborze zmiennych. Rysunek 7 wykazuje występowanie korelacji pomiędzy niektórymi zmiennymi. Szczególnie wysoką liniową zależność wykazują logarytmy zmiennych. Może to być spory problem, wrócimy do niego na etapie weryfikacji założeń.



Rysunek 7: Zależności liniowe pomiędzy zmiennymi numerycznymi

3 Budowa, kalibracja i estymacja modelu

3.1 Od ogółu do szczegółu

Przy budowie modelu zastosujemy podejście "od ogółu do szczegółu". Idea jest taka, że wychodzimy od najbardziej ogólnego modelu ze wszystkimi dostępnymi zmiennymi, a następnie redukujemy model, poprzez porównywanie ze sobą modeli zawierających odpowiednio n i $n-1$ zmiennych. Naszym kryterium decyzyjnym będzie wartość kryterium informacyjnego Akaike dla poszczególnych modeli. Zaczniemy więc od przeanalizowania modelu będącego wynikiem rozważań z rozdziału 2. Zmiennymi objaśniającymi będą zmienne uwzględnione w macierzy korelacji z podrozdziału 2.3 oraz iloczyny: gospodarstwo_2:log_koszty_p_roslinnej, gospodarstwo_2:pasze_wlasne oraz gospodarstwo_2:log_uprawy_pastewne. Będziemy nazywać go modelem nr 1. Wszystkie hipotezy będziemy testować na poziomie istotności $\alpha = 5\%$.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-2518.9434	4125.3646	-0.61	0.5430
log_naklady_pracy	-767.9925	313.0408	-2.45	0.0160
pasze_wlasne	0.7589	0.7624	1.00	0.3221
pasze_obce	4.2347	0.8657	4.89	0.0000
log_koszty_p_zwierzecej	539.1355	219.0767	2.46	0.0157
gospodarstwo_2male	4044.5284	2934.4809	1.38	0.1714
doplaty	-0.0044	0.0159	-0.28	0.7826
log_uprawy_pastewne	1335.2988	347.0917	3.85	0.0002
log_mleko	-715.4879	412.2697	-1.74	0.0860
koszty_prod_roslinnej	-2.3736	2.9155	-0.81	0.4177
log_koszty_p_roslinnej	1299.7893	660.8707	1.97	0.0522
gospodarstwo_2male:log_koszty_p_roslinnej	-226.1467	530.2323	-0.43	0.6707
pasze_wlasne:gospodarstwo_2male	-0.7120	1.0870	-0.66	0.5141
gospodarstwo_2male:log_uprawy_pastewne	-849.4424	347.8838	-2.44	0.0165

Tabela 1: Model nr 1 summary

Dla modelu nr 1 otrzymaliśmy następujące statystyki $R^2 = 0.7681$, adjusted $R^2 = 0.7357$, a dla testu F pvalue $< 2.2e-16$. Tak mała wartość pvalue oznacza, że odrzucamy hipotezę zerową o braku zależności liniowej pomiędzy zmienną objaśnianą, a zmiennymi objaśniającymi. Statystykę R^2 interpretujemy w ten sposób, że 76.81% zmienności zmiennej objaśnianej zostało wyjaśniene przez nasz model. Z kolei statystyka adjusted R^2 to statystyka R^2 z uwzględnioną karę za liczbę paramtrów. Jest jednym ze sposobów na porównywanie modeli między sobą. W modelu nr 1 wiele zmiennych objaśniających jest nieistotnych statystycznie. Stwierdzamy to na podstawie wartości pvalue dla t-testu, wykonanego dla poszczegołnych zmiennych (patrz ostat-

nia kolumna w tabeli 1). Hipoteza zerowa tego testu zakłada, że $\beta_i = 0$ (dla konkretnego i). jest to szczególny przykład testu F na układ liniowych restrykcji.

Zastosujemy teraz wspomniane wcześniej podejście "od ogółu do szczegółu", w celu redukcji modelu. Przebieg tego procesu widoczny jest w tabeli 2. Z początkowego modelu odrzuciliśmy w sumie pięć zmiennych, pozwoliło nam to zredukować kryterium AIC z 1483.87 do 1475.83. Interpretujemy to w ten sposób, że utracie tych zmiennych nie towarzyszyła znacząca utrata informacji. W tabeli poniżej prezentujemy summary dla modelu backward (tak będziemy nazywać model zbudowany metodą "od ogółu do szczegółu").

	Step	AIC
1	model nr 1	1483.87
2	- doplaty	1481.96
3	- gospodarstwo_2:log_koszty_p_roslinnej	1480.19
4	- pasze_wlasne:gospodarstwo_2	1478.61
5	- koszty_prod_roslinnej	1477.12
6	- pasze_wlasne	1475.83

Tabela 2: Kolejno odrzucane zmienne za pomocą kryterium Akaike

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-376.9677	2719.1305	-0.14	0.8900
log_naklady_pracy	-815.7388	276.6123	-2.95	0.0040
pasze_obce	4.4443	0.6952	6.39	0.0000
log_koszty_p_zwierzecej	604.2047	206.7959	2.92	0.0043
gospodarstwo_2male	3349.7337	1126.9742	2.97	0.0037
log_uprawy_pastewne	1471.2746	291.5380	5.05	0.0000
log_mleko	-846.9425	335.8121	-2.52	0.0133
log_koszty_p_roslinnej	885.7375	204.6908	4.33	0.0000
gospodarstwo_2male:log_uprawy_pastewne	-987.8298	300.9282	-3.28	0.0014

Tabela 3: Model backward

Pomimo, że adjusted R^2 nie zmienił się znacząco to udało się poprawić istotność zmiennych, co widać w tabeli 3. Jest to szczególnie ważne z punktu widzenia interpretacyjnego. Dodatkowo, po wykonaniu testu F na układ liniowych restrykcji, którego hipoteza zerowa mówi, że wartości parametrów przy zmiennych występujących w modelu nr 1, a niewystępujących w modelu backward są równe 0, otrzymujemy wysokie p-value równe 0.885, które nie daje nam podstaw do odrzucenia tej hipotezy.

Jest to dobry moment, żeby sprawdzić, czy aby na pewno użycie logarytmów poszczególnych zmiennych objaśniających było uzasadnione. Wykorzystamy więc analogiczny model do modelu nr 1, z tą różnicą, że nie będziemy obkładać żadnych zmiennych objaśniających logarytmami. Nazwiemy

go modelem prostym i również zastosujemy do niego metodę "od ogółu do szczegółu" w celu wyeliminiowania nieinformatywnych zmiennych. Tabela 4 pokazuje summary dla modelu bez logarytmów.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5451.6550	466.3600	11.69	0.0000
naklady_pracy	-184.4782	101.7818	-1.81	0.0730
pasze_wlasne	1.1457	0.5205	2.20	0.0301
pasze_obce	2.5542	0.5400	4.73	0.0000
koszty_prod_zwierzecej	0.0234	0.0074	3.18	0.0020
gospodarstwo_2male	-2886.2234	592.0493	-4.87	0.0000
uprawy_pastewne	5.9385	4.0590	1.46	0.1467
koszty_prod_roslinnej	-0.1390	1.3186	-0.11	0.9163
gospodarstwo_2male:koszty_prod_roslinnej	11.3191	3.6655	3.09	0.0026

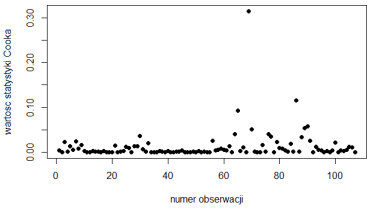
Tabela 4: Model prosty

Dla modelu prostego dostajemy adjusted $R^2 = 0.6611$, co jasno mówi nam, że użycie logarytmów było korzystnym zabiegiem. Z tego powodu dalszą analizę i badanie założeń modelu regresji liniowej będziemy przeprowadzać dla modelu backward (z logarytmami).

3.2 Identyfikacja i analiza obserwacji wpływowych

Zajmiemy się teraz identyfikacją i potencjalną eliminacją obserwacji znacząco się wyróżniających. Obserwacje tego rodzaju są niereprezentatywne, ale nierzadko mają duży wpływ na współczynniki regresji. Do zidentyfikowania obserwacji wpływowych użyjemy odległości (statystyki) Cooka. Bada ona jak zmieni się dopasowanie modelu po usunięciu jednej (i-tej) obserwacji. Wszystkie wartości tej statystyki powinny być podobnej wielkości, a jeśli nie są można przypuszczać, że dana obserwacja miała istotny wpływ na obciążenie współczynników regresji.

Na rysunku 8 występuje jedna, szczególnie wyróżniająca się obserwacja. Jest to region Campania, znajdujący się na południu Włoch. Wydajność krów wynosi tam 1608 litrów na krowę, co jest najniższą wartością w całym zbiorze danych. Po usunięciu tej obserwacji i ponownym wyestymowaniu modelu R^2 wzrosło z 0.7638 do 0.7847. Zmieniły się również oszacownia parametrów, co ilustruje tabela numer 5. Zmiany wielkości parametrów są naprawdę zauważalne. Aż sześć z dziewięciu parametrów zmieniło znacząco swoją wartość (nie liczę



Rysunek 8: Odległość Cooka dla poszczególnych obserwacji

tutaj wyrazu wolnego, reprezentowanego przez β_0, ponieważ był on statystycznie nieistotny, patrz tabela 3, ostatnia kolumna). Zmianę przy parametrze i uznajemy za znaczącą, gdy wartość tej zmiany jest podobna do wartości błędu średnego szacunku dla parametru i (patrz trzecia kolumna w tabeli 3). $\beta_0, ..., \beta_8$ to parametry odpowiadające zmiennym z modelu backward.

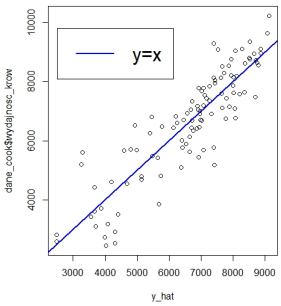
Parametr	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
model_backward	-376.96	-815.73	4.44	604.20	3349.73	1471.27	-846.94	885.73	-987.82
model_cook	-1130.94	-674.73	3.59	582.72	2295.01	1071.59	-591.36	860.19	-719.83

Tabela 5: Zmiany parametrów regresji

Przy analizie założeń, interpretacji parametrów oraz ocenie mocy prognostycznej modelu będziemy się odwoływać do parametrów z modelu z usuniętą obserwacją (modelu cooka).

4 Weryfikacja założeń

4.1 Liniowa zależność zmiennej endogenicznej



Rysunek 9: Wartości zaobserwowane vs. dopasowane

Do zbadania liniowej zależności pomiędzy zmienną objaśnianą, a zmiennymi objaśniającymi użyjemy metody graficznej, polegającej na analizie wykresu wartości zaobserwowanych względem wartości dopasowanych. Analizując rysunek 9, zauważamy, że punkty układają się symetrycznie wzdłuż prostej $y = x$, co potwierdza liniowość naszego modelu.

4.2 Współliniowość zmiennych egzogenicznych

Sprawdzimy teraz współliniowość zmiennych objaśnianych naszego modelu, gdyż odwołując się do macierzy korelacji z podrozdziału 2.3, możemy podejrzewać, że niektóre zmienne mogą powodować tę niechcianą własność. Wykorzystamy do tego celu wektor tolerancji.(indeksy w tabeli 6 to kolejne indeksy zmiennych objaśniających bez wyrazu wolnego).

Indeks	1	2	3	4	5	6	7	8
Wartość	0.413	0.282	0.076	0.021	0.082	0.045	0.493	0.037

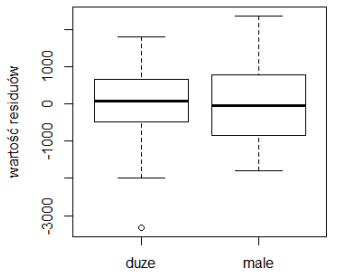
Tabela 6: Wartości wektora tolerancji dla zmiennych objaśniających

Wartości tabeli 6 nie schodzą poniżej 0.01, czyli nie zidentyfikowaliśmy, żadnej zmiennej powodującej problemy ze współliniowością. Pomimo, że niektóre wartości są dosyć niskie (wartości poniżej 0.1 mogą niepokoić), to nie odrzucamy żadnej zmiennej, powołując się na wcześniej wykonany test F, który zasugerował, że wszystkie zmienne są istotne.

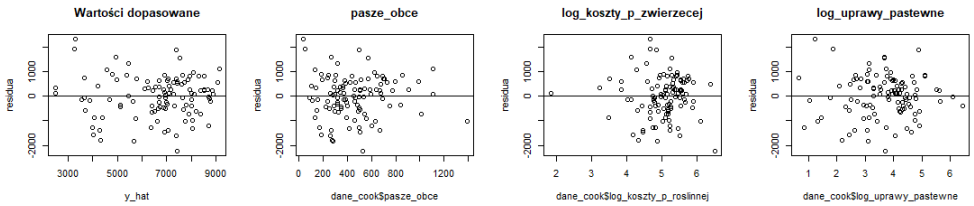
4.3 Poprawna specyfikacja błędów

Jednym z założeń regresji liniowej jest homoskedastyczność błędów. W praktyce nigdy nie mamy dostępu do informacji o błędach, możemy co najwyżej analizować residua.

Wykres 10 sugeruje, że może występować różnica pomiędzy wariancajmi pomiędzy gospodarstawmi różnych wielkośći. Hipoteza ta ma również sens ekonomiczny. Małe gospodarstwa dysponują inną technologią, środkami produkcji niż duże gospodarstwa. Aby zweryfikować tę hipotezę wykonamy test Goldfielda-Quanta, badający wariancję reszt na różnych podzbiorach. W naszym przypadku za kryterium podziału przyjmiemy wielkość gospodarstwa. Otrzymaliśmy pvalue równe 0.076, co nie daje nam podstaw do odrzucenia hipotezy zerowej postulującej brak różnic pomiędzy grupami. Jest to jednak mało konkluzywny wynik. Sprawdzimy teraz, czy czynnik losowy ε naszego modelu nie jest zależny od zmiennych objaśniających. Na rysunku 11 prezentujemy kolejne wykresy z resztami.



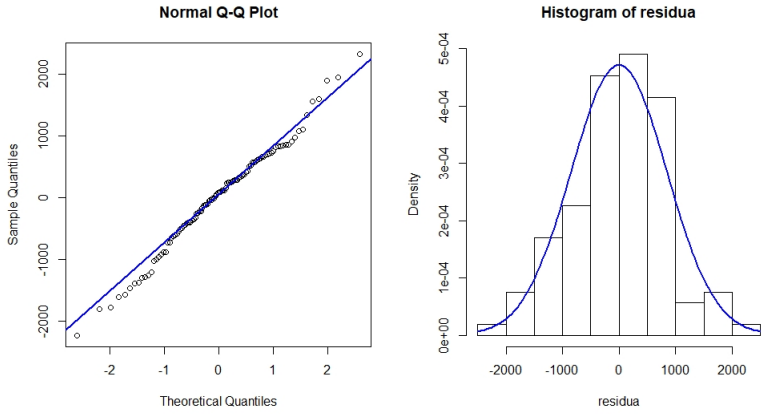
Rysunek 10: residua dla małych i dużych gospodarstw



Rysunek 11: Residua vs najistotniejsze zmienne objaśniające i y_hat

Na każdym wykresie przedstawionym na rysunku 11, reszty układają się mniej więcej symetrycznie względem prostej $y = 0$. Nie obserwujemy żadnych trendów. Aby formalnie przetestować homoskedastyczność składnika losowego wykonamy komendą bptest test Breuscha-Pagana, sprawdzający czy zmienne objaśniające mogą prognozować kwadraty residuów. Otrzymaliśmy pvalue równe 0.3136, co nie daje nam podstaw do odrzucenia hipotezy zerowej zakładającej homoskedastyczność składnika losowego.

W celu sprawdzenia normalności błędów wykorzystaliśmy test Shapiro-Wilka posługując się komendą shapiro.test. Otrzymaliśmy pvalue wynoszące 0.5389, które nie daje nam podstaw do odrzucenia H_0, mówiącej o normalności naszych residuów. Jest to zadowalający wynik, który pokrywa się z analizą graficzną wykresu kwantyl empiryczny vs. kwantyl teoretyczny rozkładu normalnego na rysunku 12. Dodatkowo kształt histogramu naszej próbki, jest zbliżony do wykresu gęstości rozkładu $\mathcal{N}(0, \sigma)$, gdzie $\sigma = 846$ to odchylenie standardowe próbki naszych residuów.



Rysunek 12: Wykres kwantylowy i histogram rozkładu residuów

5 Ocena stabilności modelu

Do zbadania stabilności modelu użyjemy technik bootstrapowych, polegających na wielokrotnym usuwaniu 20% losowych obserwacji i estymowaniu modelu dla pozostałych 80%. W naszym przypadku będzie to 1000 losowań. Wyniki doświadczenia z tabeli 7 pokazują, że wartości współczynników nie mają dużych wachań w przypadku usuwania losowych obserwacji. Średnie z tysiąca losowań są bardzo zbliżone do rzeczywistych wartości, a skrajnie oddalone wartości reprezentowane przez kwantyle rzędu 0.025 i 0.975 nie są oddalone od swoich średnich o więcej niż wartości tych średnich. Jest to

satysfakcjonujący wynik, który nie daje powodów do obaw o stabilność.

Zmienna	Wartość rzeczywista	Średnia	Odchylenie	Kwantyl rzędu 0.025	Kwantyl rzędu 0.975
log_naklady_pracy	-674.73	-663.56	135.91	-907.11	-376.22
pasze_obce	3.59	3.62	0.32	3.01	4.27
log_koszty_p_zwierzecej	582.72	575.07	141.03	273.29	830.03
gospodarstwo_2male	2295.01	2264.41	684.58	755.51	3567.24
log_uprawy_pastewne	1071.59	1069.40	149.25	777.85	1364.69
log_mleko	-591.36	-588.83	170.90	-900.41	-231.97
log_koszty_p_roslinnej	860.19	872.49	108.98	698.73	1120.27
gospodarstwo_2male:log_uprawy_pastewne	-719.83	-710.96	201.76	-1092.54	-253.79

Tabela 7: Zachowanie modelu w przypadku usuwania obserwacji

6 Moc prognostyczna modelu

Do oceny mocy prognostycznej modelu wykorzystamy metodę zwaną sprawdzianem krzyżowym (ang. cross-validation). Pierwszy krok polega na podzieleniu zbioru danych na k podzbiorów o podobnej liczności. Następnie po kolei wybieramy każdy z k podzbiorów, "odkładamy" go na bok, uczymy model na pozostałych $k-1$ podzbiorach i testujemy na odłożonym zbiorze. Całą tą procedurę wykonujemy k razy. Wybór wielkości parametru k jest subiektywny, więc decydujemy się na $k = 6$, dodatkowo całą procedurę sprawdzianu krzyżowego powtórzymy trzy razy. W sumie dla 18 próbek średnia wartość R^2 wyniosła 0.759, z odchyleniem standardowym równym 0.095. W tym kontkeście odchylenie standardowe interpretujemy jako miarę niepewności, jego wartość 0.095 oznacza występowanie istotnych różnic w wartościach statystyki R^2 pomiędzy poszczególnymi próbami. Ważną kwestią jest tutaj porównanie średniego wyniku R^2 dla CV (0.759) z R^2 dla naszego finalnego modelu ($R^2 = 0.785$). Różnica jest niewielka, co oznacza, że nasz model nie jest przetrenowany i bez obaw możemy używać go w celach prognostycznych.

7 LASSO

Regresja metodą LASSO charakteryzuje się wprowadzeniem dodatkowego czynnika, który "karze" parametry za ich wielkość. Zmienia to postać funkcji kosztu (straty) dla modelu. Przy korzystaniu z regresji metodą LASSO rekomendowane jest, aby zmienne były na podobnej skali. W naszym przypadku warunek ten nie jest spełniowy, więc dane poddamy standaryzacji (a potem wykonamy odwrotną procedurę aby móc porównywać paramtery). Naszym zestawem zmiennych będzie wyjściowy zestaw zmiennych dla modelu nr 1. Istotność czynnika karzącego wyznacza parametr λ. Przymujemy $\lambda = 10$ (wyboru dokonaliśmy na podstawie przeprowadzenia sprawdzianiu krzyżowego dla 40 różnych wartości λ). Tabela 8 pokazuje różnice pomiędzy bada-

nymi modelami. Dla części zmiennych, w modelu LASSO (gospodarstwo2, log_mleko i doplaty) współczynnik równy jest zero. Jest to odpowiednikiem usunięcia zmiennej.

Parametry przy zmiennych pojawiających się w obydwu modelach mają podobne wielkości. W regresji metodą LASSO ciekawą kwestią jest wyzerowanie się współczynnika przy zmiennej gospodarstwo2 i jednoczesna obecność wszystkich trzech iloczynów. Wygląda to na alternatywny sposób uwzględnienia wpływu wielkości gospodarstwa. Na zakończenie tego rozdziału porównamy jeszcze zdolności prognostyczne obu modeli. W tym celu zastosujemy analogiczne podejście jak w rodziale 6. Dla regresji metodą LASSO otrzymujemy średni R^2 równy 0.74 z odchyleniem standardowym równym 0.126. Nasz finalny model otrzymał tutaj wyniki 0.759 i 0.095, które są nieznacznie lepsze od wyników otrzymanych metodą LASSO.

zmienna	LASSO	model finalny
intercept	-3081.336	-1130.94
log_naklady_pracy	-560.28	-674.73
pasze_wlasne	1.00	brak
pasze_obce	2.18	3.59
log_koszty_p_zwierzecej	453.57	582.72
doplaty	0.00	brak
log_uprawy_pastewne	492.17	1071.59
log_mleko	0.00	-591.36
log_koszty_p_roslinnej	630.34	860.19
gospodarstwo_2	0.00	2295.01
iloczyn_koszty_p_zwierzecej	182.80	brak
iloczyn_pasze_wlasne	-0.57	brak
iloczyn_log_uprawy_pastewne	-304.56	-719.83

Tabela 8: Regresja LASSO

8 Interpretacja współczynników modelu

Nasz ostateczny model przyjął postać:

$$y = \beta_0 + \beta_1 lnx_1 + \beta_2 x_2 + \beta_3 lnx_3 + \beta_4 x_4 + \beta_5 lnx_5 + \beta_6 lnx_6 + \beta_7 lnx_7 + \beta_8 x_4 lnx_5$$

Opis poszczególnych zmiennych znajduje się w tabeli 9.

Wpływ każdej ze zmiennych interpretujemy osobno, przy założeniu ceteris paribus. Parametr β_1 interpretujemy tak, że gdy nakłady pracy w gospodarstwach rosną o 1% to wydajność krów spada o $\frac{674.73}{100}$ kg mleka na krowę (korzystamy z przybliżenia $ln(x+1) \approx x$ dla małych wartości x). Są do zaskakujące wnioski, sprzeczne z teorią ekonomii. Interpretacją ekonomiczną mogłoby być tutaj stwierdzenie, że większe nakłady pracy nie zwiększają wydajności krów. Krowy prawdopodobnie nie

y	wydajność krów
x_1	nakłady pracy
x_2	pasze obce
x_3	koszty prod. zwierzęcej
x_4	gospodarstwo (duże / małe)
x_5	uprawy pastewne
x_6	produkcja mleka
x_7	koszty prod. roślinnej

Tabela 9: Opis zmiennych

potrzebują dużo doglądu przez człowieka, ważniejsze są w ich przypadku inne czynniki. Wartość 258.67, czyli średni błąd szacunku dla parametru β_1 interpetujemy tak, że oszacowany parametr może się średnio różnić od jego rzeczywistej wartości o ± 258.66. Parametr β_2 interpretujemy wprost, że wraz ze wzrostem inwestycji w pasze obce o jedną złotówkę na krowę, jej wydajność wzrośnie o 3.59 kg mleka na rok. Może być to interpretowane jako zakup wysokojakościowych produktów, które według modelu pozytywnie wpływają na funkcjonowanie tych zwierząt. Jest to jedyna niezmodyfikowana zmienna istotna statystycznie, o najwyższej wartości statystyki w t-teście, co oznacza, że jest to najbardziej istotna zmienna występująca w naszym modelu. Parametr β_3 może być interpretowany w analogiczny sposób jak β_1, z tą różnicą, że tym razem wzrost wydajności krów o $\frac{582.72}{100}$kg, pod wpływem wzrostu kosztów produkcji zwierzęcej o 1% wydaje się być czymś naturalnym i zgodnym ze zdrowym rozsądkiem. Zwierzęta zadbane, zdrowe i posiadające odpowiednią opiekę weterynaryjną są bardziej wydajne.

Parametr	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
model_cook	-1130.94	-674.73	3.59	582.72	2295.01	1071.59	-591.36	860.19	-719.83
bł. śr. szacunku	2527.10	258.67	0.67	191.76	1075.41	287.04	317.37	189.83	286.39

Tabela 10: Model finalny

Aby rozważyć wpływ wielkości gospodarstwa na wydajność krów, policzymy pochodną cząstkową z y po x_4, z czego dostajemy $\frac{\partial y}{\partial x_4} = \beta_4 + \beta_8 ln x_5$. Widzimy więc, że aby ocenić rzeczywisty wzrost wydajności w tym przypadku, musimy posiadać informację na temat aktualnej wartości zmiennej x_5 (uprawy pastewne, w hektarach). Przyjmując, że $lnx_5 = z$, obserwujemy zmianę wydajności krów mlecznych na mniejszych gospodarstwach o $(2295.01 - 719.83z)$ jednostek. Badanie wpływu wielkości upraw również będzie od nas wymagało informacji o wielkości gospodarstwa. Wraz z powiększeniem się powierzchni upraw pastewnych o 1% wydajność krów wzrośnie o $\frac{2295.01-719.83}{100}$ kg dla gospodarstw małych lub o $\frac{2295.01}{100}$ kg dla gospodarstw dużych. Widzimy więc, że zwiększanie areału pod uprawy jest bardziej opłacalne na dużych gospodarstwach. Jeśli rozważymy teraz wpływ ilości produkowanego mleka, to zauważymy, że podobnie jak w przypadku zmiennej x_1, wraz ze wzrostem wartości tej zmiennej o 1% wydajność krów spadnie o $\frac{591.36}{100}$ kg mleka rocznie. Prawdopodobnie wynika to stąd, że gospodarstwa produkujące duże ilości mleka nie są w stanie maksymalnie efektywnie wykorzystać potencjału swojego bydła, przez masową produkcję. Wzrost kosztów produkcji roślinnej (w zł/ha) o 1% doprowadzi do wzrostu wydajności krów o $\frac{860.19}{100}$ kg. Zwiększanie jakości pastwisk naturalnie może wpływać pozytywnie na wypasane na nim krowy. Ostatnią ważną kwestią jest wyraz wolny β_0, któ-

ry w naszym modelu jest nieinterpretowalny, gdyż zazwyczaj utożsamia się go z wartością oczekiwaną zmiennej endogenicznej w przypadku zerowania się wartości zmiennych egzogenicznych, co w naszym przypadku nie miałoby sensu z ekonomicznego punktu widzenia.

9 Podsumowanie

Badając wpływ czynników zestawionych w rozdziale 1 na wydajność europejskich krów mlecznych doprowadziliśmy do utworzenia modelu, który spełnia wszystkie założenia dla klasycznego normalnego modelu regresji liniowej. Dzięki temu, byliśmy w stanie wyciągać wnioski ze wszystkich testów statystycznych (m.in. F-testy i t-testy), które miały kluczowe znaczenie w momentach decyzyjnych o odrzucaniu lub zmodyfikowaniu konkretnych modeli lub zmiennych (rozdział 3). Dużą rolę w trakcie budowy modelu pełniły też obserwacje wpływowe, które miały znaczny wpływ na jego postać. Ostatecznie model okazał się być odporny na zmiany lub ubytki obserwacji, a sprawdzian krzyżowy wykazał też, że nie jest on przetrenowany. Udało się nam uzyskać lepszy poziom dopasowania niż dla regresji metodą Lasso. Wpływ zmiennych istotnych w większości przypadków okazał się racjonalny, ale też pojawiły się dosyć specyficzne własności, których spodziewaliśmy się rozpoczynając pracę z naszym zestawem danych. Wszystkie powyższe czynniki skłoniły nas do przyjęcia modelu finalnego.