AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Ekonometria

Projekt 5

Temat: Wstępna analiza danych służących do stworzenia modelu regresji wielorakiej — część druga

Autorzy: Laura Cygan, Wiktoria Bąk

Wydział: Wydział Zarządzania

Kierunek: Informatyka i Ekonometria

Przedmiot: Ekonometria

Kraków, 2024

WSTĘP

W drugiej części wstępnej analizy danych, które następnie zostaną wykorzystane przez nas do stworzenia modelu regresji wielorakiej, skupiliśmy się na analizie obserwacji odstających, aby zadecydować, które z nich odrzucić. Sprawdziliśmy również, które ze zmiennych objaśniających wybrać i czy stworzony w ten sposób model spełnia założenia Gaussa-Markowa. Wszystkie części projektu wykonaliśmy dla trzech zestawów danych – uwzględniając wszystkie powiaty, a także stosując podział na te grodzkie i ziemskie.

Dla przypomnienia, jako zmienną objaśnianą do naszego modelu wybraliśmy bezrobocie (w %). W celu próby wyjaśnienia, co wpływa na procent niezatrudnionych ludzi w poszczególnych powiatach, wybraliśmy pięć potencjalnych zmiennych objaśniających: średnie wynagrodzenie, współczynnik feminizacji, współczynnik urbanizacji, liczbę ofert pracy i liczbę małżeństw przypadających na 10 tys. mieszkańców.

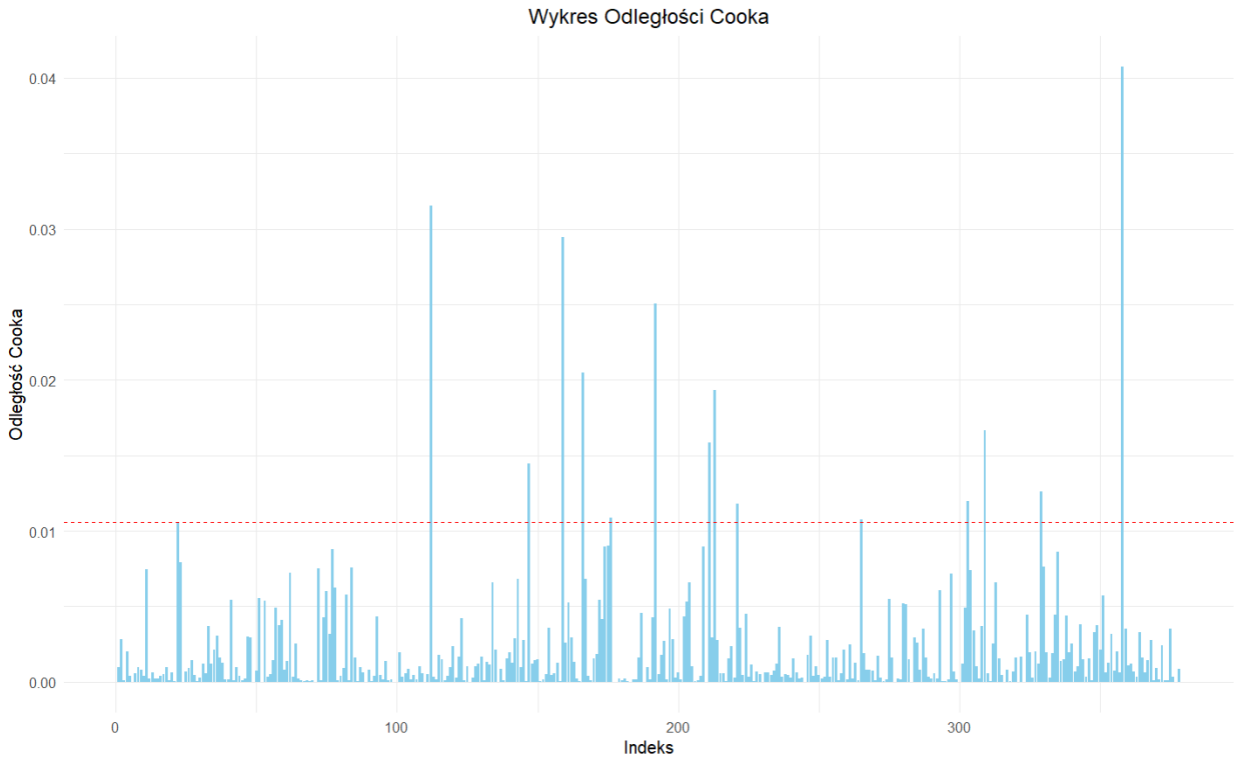
CZĘŚĆ PROJEKTOWA

Dla wszystkich powiatów

Analiza występowania wartości odstających, dźwigniowych i wpływowych

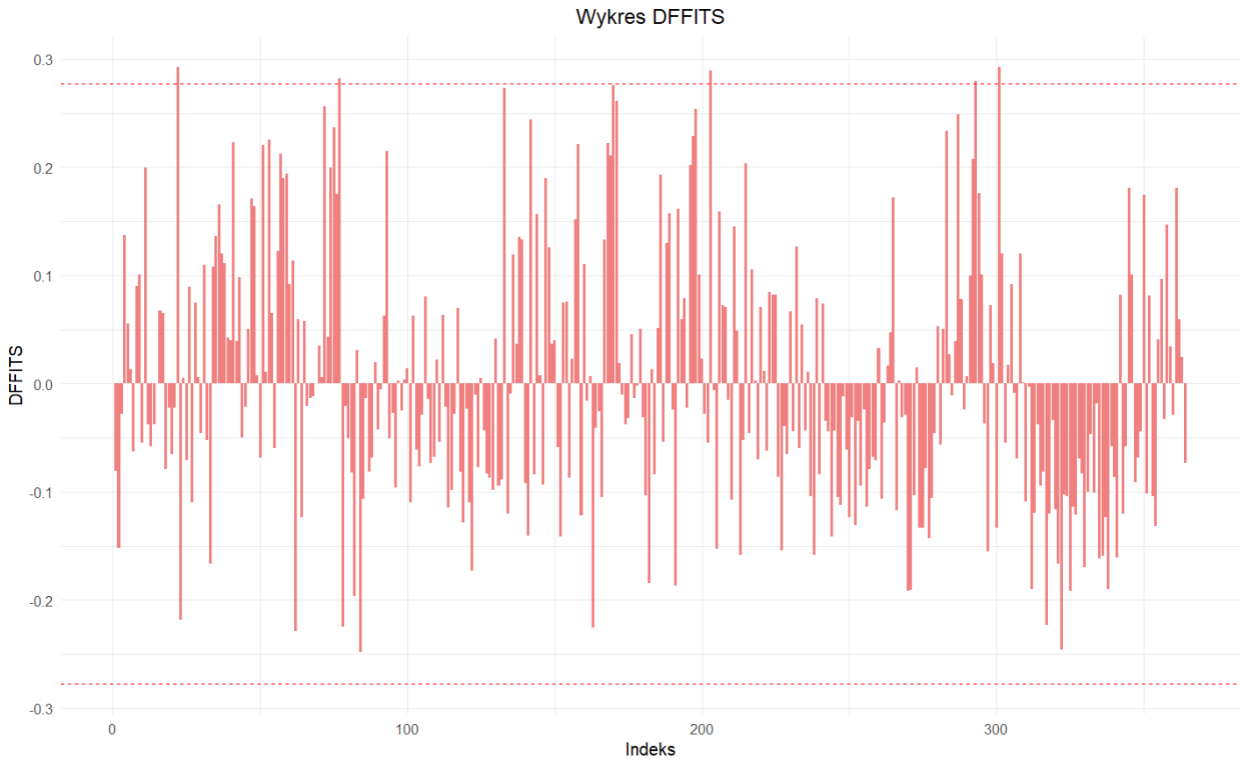
Pobrane przez nas dane zawierają informacje o 378 powiatach w Polsce. Ogółem jest ich 380, jednak dla dwóch Bank Danych Lokalnych GUS nie posiadał potrzebnych nam informacji.

W celu zidentyfikowania wartości wpływowych obliczyliśmy dla wszystkich obserwacji odległość Cooka. Za wartość progową uznaliśmy $\frac{4}{n}$, czyli w naszym przypadku 0.01.



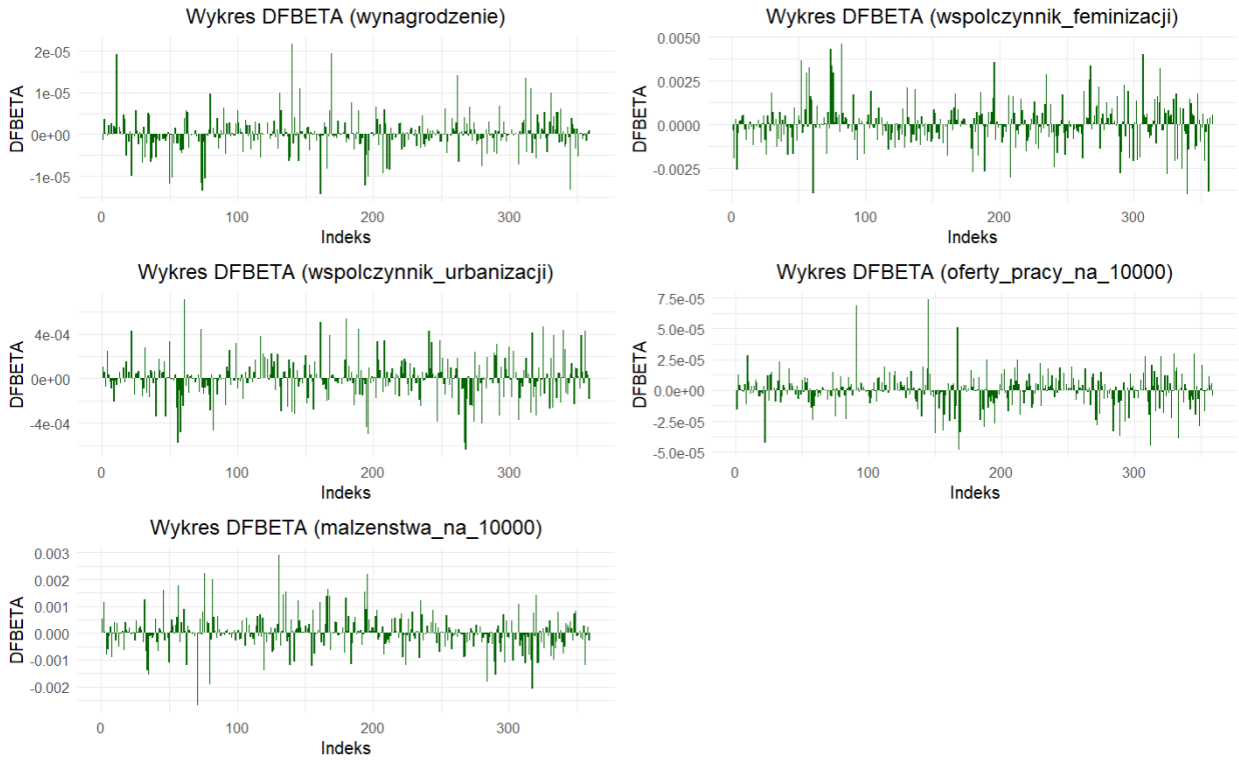
Zgodnie w powyższym wykresem 14 obserwacji ma odległość Cooka większą niż przyjęta przez nas wartość progowa. Odrzucamy te wartości, w związku z czym zostają nam dane dla 364 powiatów.

Następnie, dla wszystkich pozostałych obserwacji wyznaczyliśmy DFFITS (Difference in FITs), aby określić wpływ i- tej obserwacji na parametry modelu. Jako wartość progową uznaliśmy $2 \cdot \sqrt{\left(\frac{k`}{n}\right)}$, czyli w naszym przypadku 0.277.



Pięć obserwacji ma wartość DFFITS większą niż przyjęty przez nas próg, w związku z czym je usuwamy.

Obliczamy DFBETA (Difference in Betas), aby zmierzyć wpływ i-tej obserwacji na oszacowanie poszczególnych parametrów modelu. Za wartość progową przyjęliśmy $\frac{2}{\sqrt{n}}$, czyli w naszym przypadku 0.105.



Jak można zaobserwować z powyższych wykresów, DFBETA nie wykrywa żadnej obserwacji jako wpływowej. Zostajemy więc z danymi dla 359 powiatów, po usunięciu w sumie 19 obserwacji. W poprzedniej części projektu (Projekt 4) zdecydowaliśmy się nie redukować żadnych danych, gdyż bazowaliśmy jedynie na statystykach opisowych i wykresach pudełkowych zmiennych. Powyższe miary jednakże wyraźnie wskazały wartości odstające i wpływające na model, które warto usunąć.

Analiza doboru zmiennych objaśniających

W tworzeniu modelu regresji wielorakiej kluczowy jest dobór odpowiednich zmiennych objaśniających. W poprzedniej części (Projekt 4) wykorzystaliśmy metodę Hellwiga, która polega na wyborze zestawu zmiennych o największej pojemności informacyjnej. Zinterpretowaliśmy również korelację między zmiennymi.

Zmienne objaśniające nie powinny być zbyt silnie ze sobą skorelowane, gdyż może to spowodować problem współliniowości, co skutkuje wzrostem niepewności odnośnie oszacowań wartości parametrów. Jednym ze sposobów sprawdzania, czy w modelu występuje współliniowość jest VIF (Variance Inflation Ratio), czyli czynnik inflacji wariancji. Gdy VIF dla danej zmiennej jest małe, to nie mamy do czynienia z problemem współliniowości.

Wynagrodzenie	Współczynnik feminizacji	Współczynnik urbanizacji	Oferty pracy	Liczba małżeństw
1.352333	3.289353	3.191689	1.093580	1.047815

Jak widać, w naszym modelu VIF dla poszczególnych zmiennych objaśniających jest raczej niewielki (VIF < 10). Oznacza to, że współliniowość nie występuje w modelu.

Aby wybrać najodpowiedniejszy zestaw zmiennych objaśniających, zastosowaliśmy metodę regresji krokowej. Polega ona na usuwaniu ze zbioru potencjalnych zmiennych objaśniających zmiennych, które nie wnoszą istotnego wkładu do opisu zmiennej objaśnianej. Wykonujemy dwa warianty: regresję krokową opartą na statystyce F oraz wykorzystując AIC, czyli kryterium informacyjne Akaike.

W RStudio można zobaczyć, jak dokładnie przebiegał proces wyboru zmiennych objaśniających przez program. Oba warianty (z kryterium informacyjnym i ze statystyką F) jako zmienne objaśniające wybrały wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy oraz liczbę małżeństw. Odrzucony został więc **współczynnik urbanizacji**, co zgadza się ze wnioskami z poprzedniego projektu (uzyskanymi za pomocą metody Hellwiga).

Weryfikacja założeń twierdzenia Gaussa-Markowa

Następnie, mając już wybrane obserwacje i odpowiednie zmienne objaśniające do modelu należy zweryfikować, czy prezentowane dane spełniają założenia twierdzenia Gaussa-Markowa.

Założenia:

1. Elementy macierzy $\boldsymbol{X}$ są nielosowe.
2. $r(\boldsymbol{X}) \geq k^*$ (tym samym $n \geq k^*$),
3. dla każdego i: $E(\varepsilon_i) = 0$, tzn. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$,
4. dla każdego i: $D^2(\varepsilon_i) = \sigma^2$ (homoskedastyczność),
5. $cov(\varepsilon_i, \varepsilon_j) = 0$ dla $i \neq j$ (brak autokorelacji)

6*. Składniki losowe ε_i mają rozkład normalny.

Twierdzenie Gaussa-Markowa

Jeżeli spełnione są założenia 1-5, to estymatory MNK są najlepszymi nieobciążonymi estymatorami w klasie estymatorów liniowych (BLUE – Best Linear Unbiased Estimators).

Homoskedastyczność

Aby sprawdzić homoskedastyczność w modelu, czyli stałość wariancji składnika losowego dla różnych wartości zmiennych niezależnych przeprowadziliśmy test Breuscha-Pagana.

P-value: 0.0002228107

Wartość mniejsza od α = 0.05 sugeruje odrzucenie hipotezy zerowej, co oznacza, że w modelu występuje **heteroskedastyczność**.

Autokorelacja składników losowych

Nie rozważamy problemu autokorelacji, gdyż dotyczy on przypadków, gdy zmienne w modelu zależą od czasu (tzn. gdy są to szeregi czasowe).

Liniowość modelu

Aby zweryfikować liniowość modelu, przeprowadziliśmy dwa testy – test RESET i test Rainbowa.

P-value dla testu RESET: 0.1116523

P-value dla testu Rainbowa: 0.08932116

W obu przypadkach wartość p-value przekracza poziom $\alpha = 0.05$, co oznacza, że nie ma podstaw do odrzucenia hipotez zerowych. Wskazuje to na **liniowość** sprawdzanego modelu.

Normalność składników losowych

Aby sprawdzić, czy składniki losowe modelu posiadają rozkład normalny, przeprowadziliśmy test Shapiro-Wilka.

P-value: 3.26338e-07

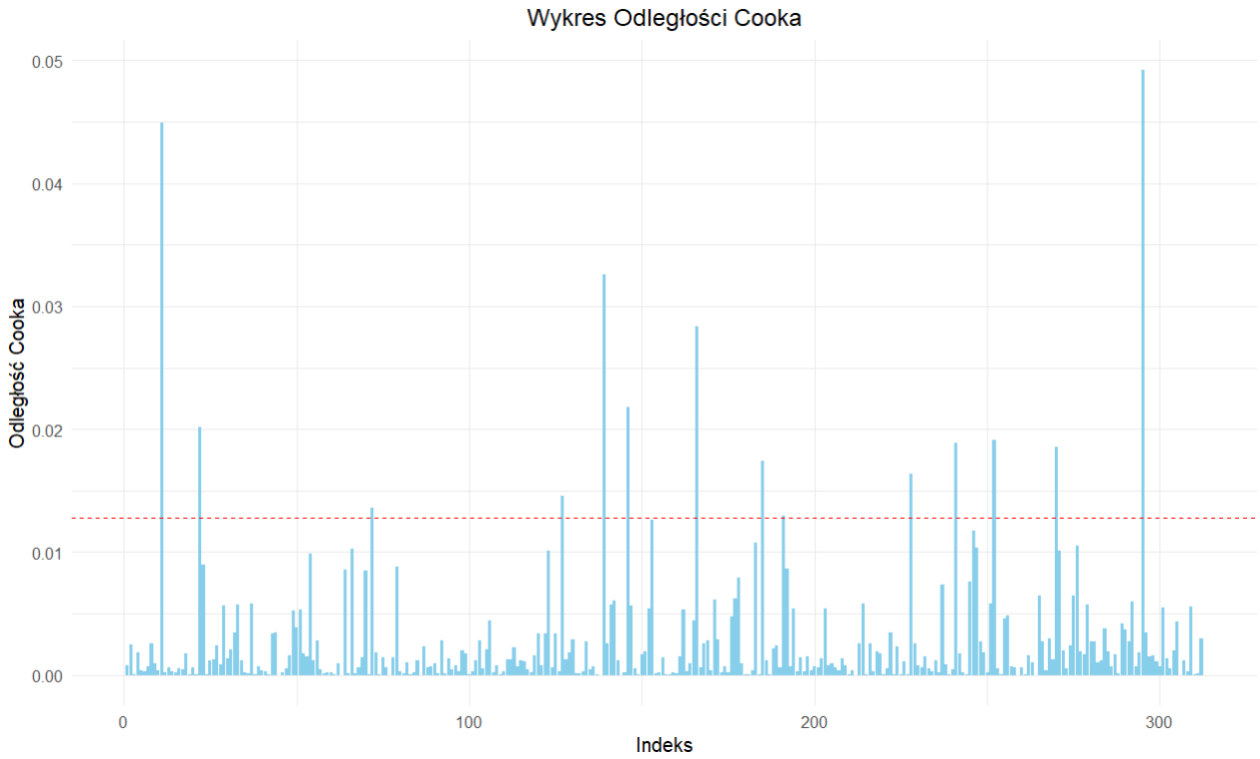
Wartość p-value jest znacznie mniejsza od $\alpha = 0.05$, co sugeruje odrzucenie hipotezy zerowej. Składniki losowe modelu **nie posiadają więc rozkładu normalnego.**

Niespełnione jest więc założenie o homoskedastyczności oraz normalności składników losowych. Może to prowadzić m.in. do błędnych decyzji dotyczących istotności zmiennych w modelu. Aby zlikwidować ten problem, można spróbować zmodyfikować zmienne, np. poprzez ich zlogarytmowanie.

Dla powiatów ziemskich

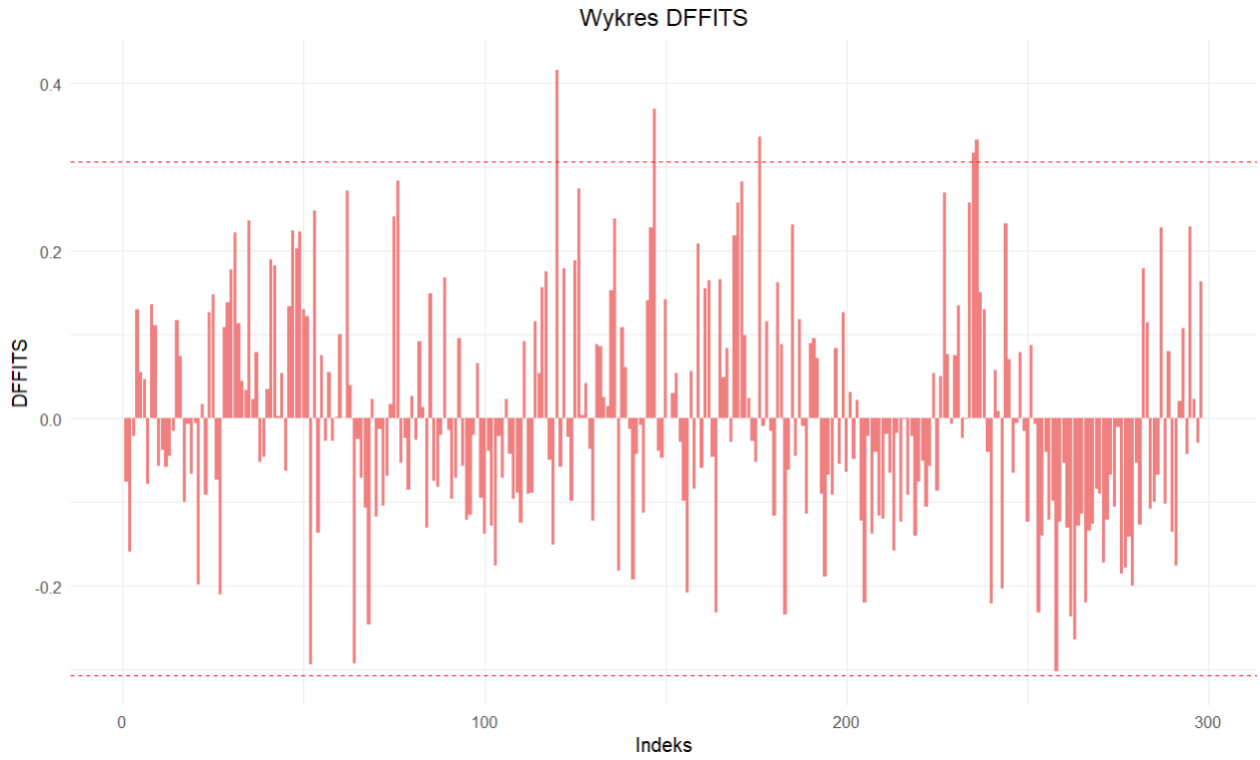
Analiza występowania wartości odstających, dźwigniowych i wpływowych

Wybierając dane jedynie dla powiatów ziemskich, dostajemy 312 obserwacji (z 314 powiatów, dwa nie posiadały danych). Ponownie, aby zidentyfikować obserwacje wpływowe, obliczamy odległości Cooka dla wszystkich danych. Przyjęta wartość progowa to 0.012.



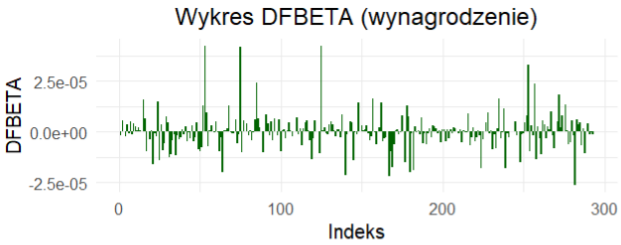
Wartość progową przekraczało 14 obserwacji, które usuwamy z danych.

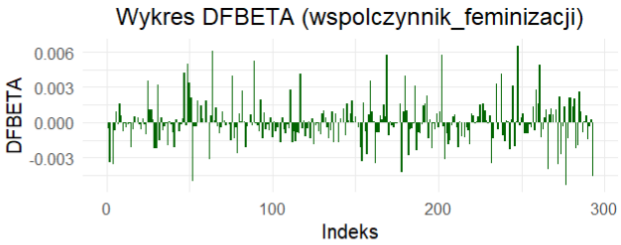
Następnie wyznaczamy DFFITS. Jako wartość progową bierzemy 0.306.

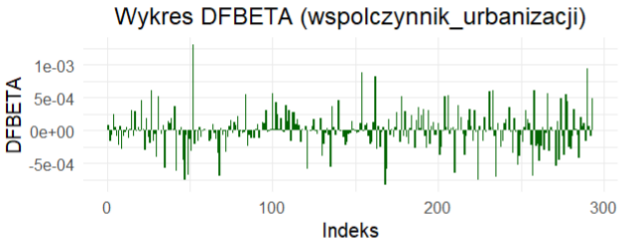


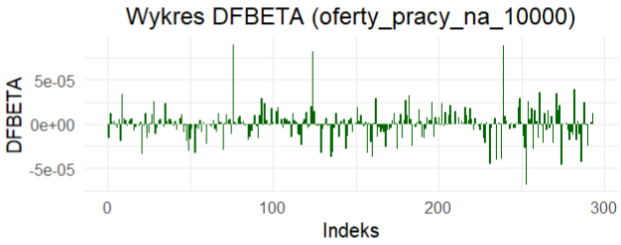
5 obserwacji przekracza wyznaczony próg. Po ich usunięciu zostają nam dane dla 293 powiatów ziemskich.

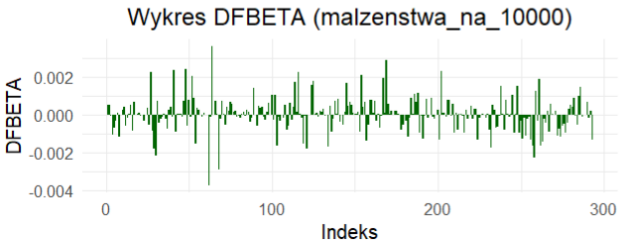
Wyznaczamy również miarę DFBETA dla każdej ze zmiennych objaśniających. Wartość progowa to 0.116.











Jak widać z wykresów, dla żadnej zmiennej obserwacje nie przekraczają wartości progowej. Nie zmieniamy więc ustalonego wcześniej zestawu danych.

W procesie identyfikowania wartości odstających usunęliśmy w sumie 19 obserwacji.

Analiza doboru zmiennych objaśniających

W celu sprawdzenia współliniowości obliczyliśmy VIF.

Wynagrodzenie	Współczynnik feminizacji	Współczynnik urbanizacji	Oferty pracy	Liczba małżeństw
1.149639	1.674914	1.596889	1.082159	1.166371

Dla wszystkich zmiennych wartości VIF oscylują na poziomie 1-2, co jest wystarczająco małą wartością, by sugerować brak współliniowości.

Następnie wykonaliśmy regresję krokową, zarówno opartą na kryterium informacyjnym Akaike, jak i statystyce F. Oba warianty jako zmienne objaśniające wybrały wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy oraz liczbę małżeństw, odrzucając **współczynnik urbanizacji**. Jest to zgodne z zestawem zmiennych wyznaczonych metodą Hellwiga w poprzednim projekcie (Projekt 4).

Weryfikacja założeń twierdzenia Gaussa-Markowa

Mając wybrane odpowiednie dane i zmienne objaśniające do modelu sprawdziliśmy, czy spełniają one założenia przytoczonego wyżej twierdzenia Gaussa-Markowa.

Homoskedastyczność

Aby sprawdzić homoskedastyczność w modelu, przeprowadziliśmy test Breuscha-Pagana.

P-value: 0.01863643

Wartość mniejsza od $\alpha = 0.05$ sugeruje odrzucenie hipotezy zerowej, co oznacza, że w modelu występuje **heteroskedastyczność**.

Autokorelacja składników losowych

Ponownie **nie rozważamy problemu autokorelacji**, gdyż dotyczy on przypadków, gdy zmienne w modelu zależą od czasu (tzn. gdy są to szeregi czasowe).

Liniowość modelu

Aby zweryfikować liniowość modelu, przeprowadziliśmy dwa testy – test RESET i test Rainbowa.

P-value dla testu RESET: 0.1753136

P-value dla testu Rainbowa: 0.1632945

W obu przypadkach wartość p-value przekracza poziom $\alpha = 0.05$, co oznacza, że nie ma podstaw do odrzucenia hipotez zerowych. Wskazuje to na **liniowość** sprawdzanego modelu.

Normalność składników losowych

Aby sprawdzić, czy składniki losowe modelu posiadają rozkład normalny, przeprowadziliśmy test Shapiro-Wilka.

P-value: 5.775385e-06

Wartość p-value jest znacznie mniejsza od $\alpha = 0.05$, co sugeruje odrzucenie hipotezy zerowej. Składniki losowe modelu **nie posiadają więc rozkładu normalnego.**

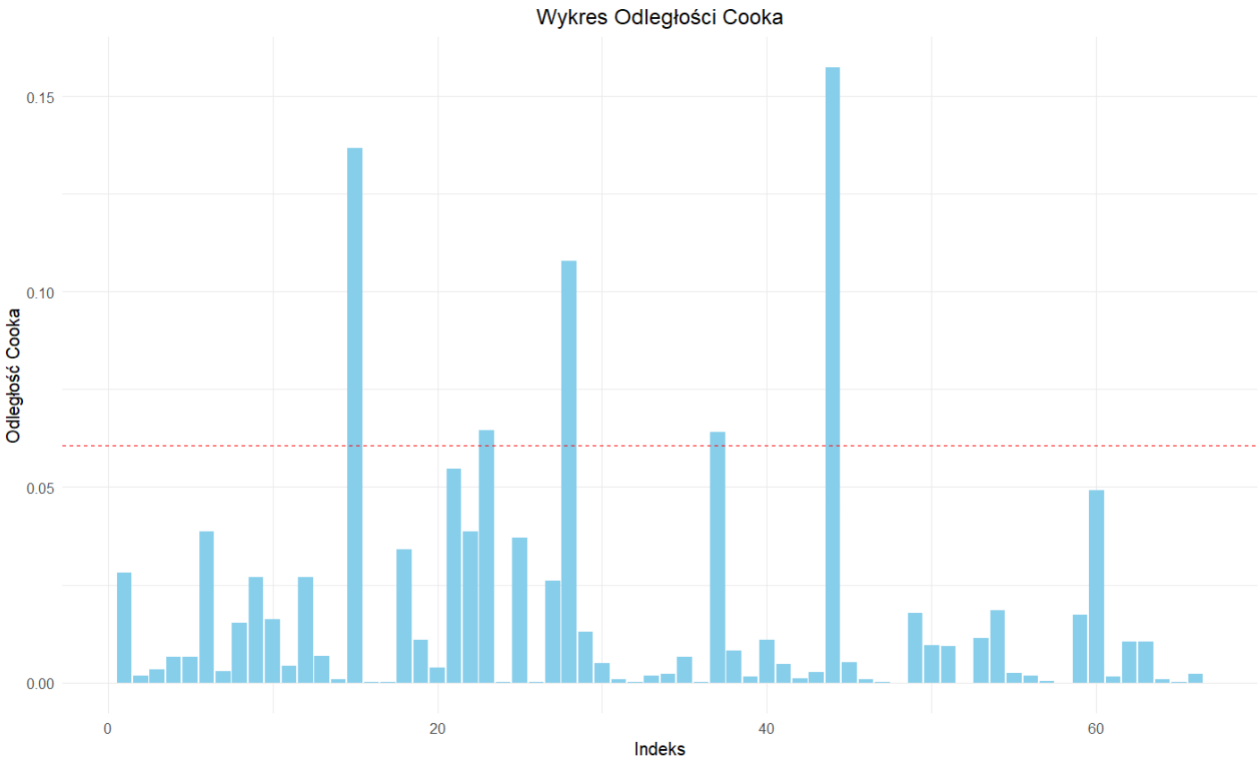
Tak samo jak w przypadku uwzględnienia danych dla wszystkich powiatów, niespełnione są dwa założenia: założenie o homoskedastyczności oraz dodatkowe o normalności składników losowych. Aby zlikwidować ten problem, można spróbować zmodyfikować zmienne, przykładowo poprzez ich zlogarytmowanie.

Dla powiatów grodzkich

Analiza występowania wartości odstających, dźwigniowych i wpływowych

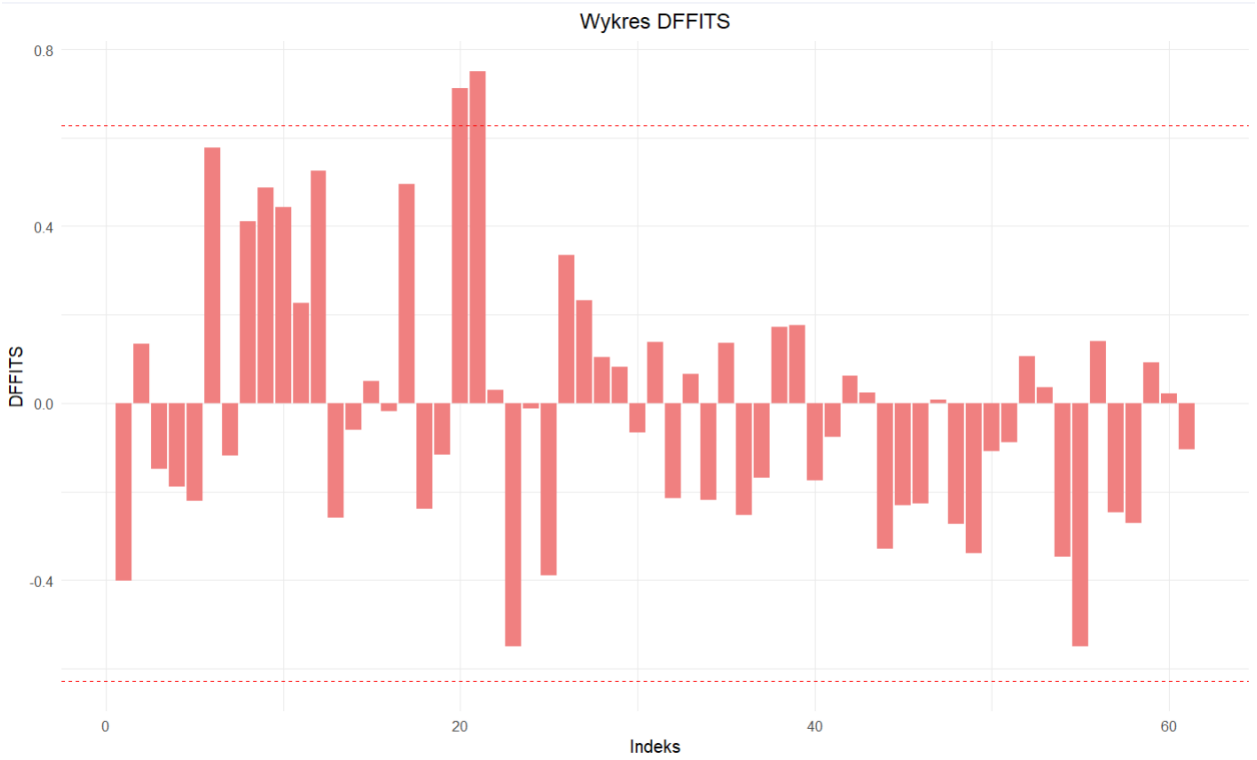
W trzecim przypadku rozważamy dane ze wszystkich 66 miast na prawach powiatu. Z tego powodu w modelu początkowym nie uwzględniamy współczynnika urbanizacji – dla każdego powiatu wynosi on 100%.

Aby zidentyfikować obserwacje wpływowe, liczymy odległości Cooka. Jako wartość progową przyjmujemy 0.06.



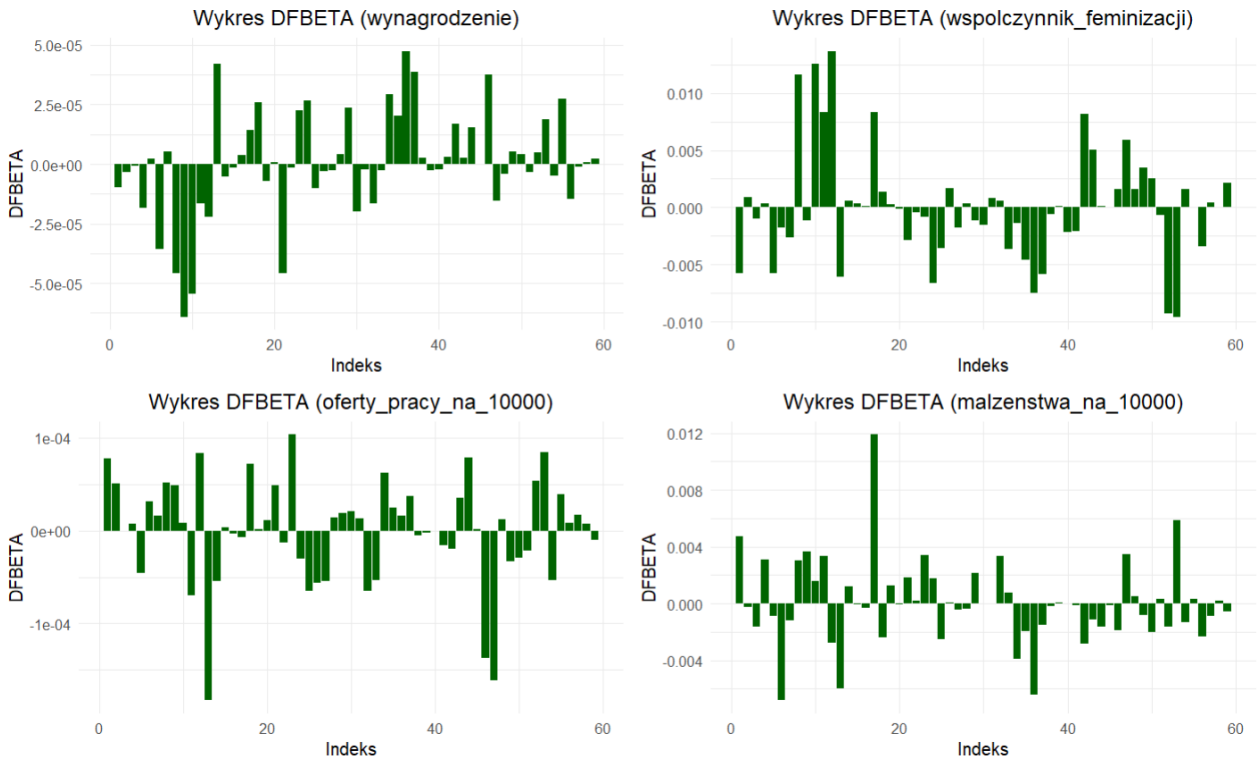
Pięć obserwacji przekracza wyznaczony przez nas próg. Usuwamy je więc ze zbioru danych.

Następnie wyznaczamy miarę DFFITS. Za wartość progową przyjmujemy 0.627.



Na wykresie obserwujemy 2 wartości wpływowe. Po ich usunięciu zostają nam dane dla 59 powiatów grodzkich.

Następnie wyznaczamy DFBETA dla każdej zmiennej objaśniającej. Wartość progowa to 0.26.



Jak widać z wykresów, żadna z obserwacji nie przekracza wyznaczonego przez nas progu.

Z 66 obserwacji ostatecznie usunęliśmy 7, zostawiając dane dla 59 powiatów.

Analiza doboru zmiennych objaśniających

W celu sprawdzenia współliniowości obliczyliśmy VIF.

Wynagrodzenie	Współczynnik feminizacji	Oferty pracy	Liczba małżeństw
1.463791	1.182169	1.010222	1.418432

Dla wszystkich zmiennych wartości VIF są względnie małe i sugerują brak współliniowości.

Następnie wykonaliśmy regresje krokowe oparte na kryterium informacyjnym Akaike i statystyce F. Oba warianty jako zmienne objaśniające wybrały wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy oraz liczbę małżeństw, **nie odrzucając żadnej ze zmiennych**. Jest to zgodne z zestawem zmiennych wyznaczonych metodą Hellwiga w poprzednim projekcie (Projekt 4).

Weryfikacja założeń twierdzenia Gaussa-Markowa

Mając wybrane odpowiednie dane i zmienne objaśniające do modelu sprawdziliśmy, czy spełniają one założenia przytoczonego wyżej twierdzenia Gaussa-Markowa.

Homoskedastyczność

Aby sprawdzić homoskedastyczność w modelu, przeprowadziliśmy test Breuscha-Pagana.

P-value: 0.1055539

Wartość większa od $\alpha = 0.05$ oznacza brak podstaw do odrzucenia hipotezy zerowej, czyli w modelu występuje **homoskedastyczność**.

Autokorelacja składników losowych

Ponownie **nie rozważamy problemu autokorelacji**, gdyż dotyczy on przypadków, gdy zmienne w modelu zależą od czasu (tzn. gdy są to szeregi czasowe).

Liniowość modelu

Aby zweryfikować liniowość modelu, przeprowadziliśmy dwa testy – test RESET i test Rainbowa.

P-value dla testu RESET: 0.02389493

P-value dla testu Rainbowa: 0.05392377

W przypadku testu Rainbowa p-value nieznacznie przekracza poziom $\alpha = 0.05$, natomiast dla testu RESET p-value wynosi jedynie 0.02, co sugeruje odrzucenie hipotezy zerowej o liniowości modelu. **Wyniki testów są więc niejednoznaczne**, szczególnie gdybyśmy założyli inny poziom istotności, jak np. $\alpha = 0.01$ lub $\alpha = 0.1$.

Normalność składników losowych

Aby sprawdzić, czy składniki losowe modelu posiadają rozkład normalny, przeprowadziliśmy test Shapiro-Wilka.

P-value: 0.06642529

Wartość p-value jest nieznacznie większa od $\alpha = 0.05$, co sugeruje brak podstaw do odrzucenia hipotezy zerowej. Składniki losowe modelu **mają więc rozkład normalny**.

W przypadku modelu dla powiatów grodzkich spełnione są wszystkie założenia twierdzenia Gaussa-Markowa z wyjątkiem liniowości modelu, której testy statystyczne wyszły niejednoznaczne. Modyfikowanie danych nie wydaje się konieczne, ale można spróbować np. zlogarytmować je jak w pozostałych przypadkach i sprawdzić, czy rozwiązuje to problem niejednoznacznej liniowości.

Podsumowanie

W każdym przypadku za pomocą odległości Cooka, miary DFFITS i DFBETA usunęliśmy obserwacje wpływowe, które mogły zaburzać modele. Miara VIF wykazała brak współliniowości. Za pomocą regresji krokowej wyznaczyliśmy ostateczny zestaw zmiennych objaśniających: **wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy oraz liczbę małżeństw.** Jest on taki sam dla każdej z grup powiatów. W przypadku wszystkich powiatów oraz powiatów ziemskich niespełnione są założenia twierdzenia Gaussa-Markowa o homoskedastyczności i normalności składników losowych. Model dla powiatów grodzkich spełnia te założenia, jednakże niejednoznacznie wyszły testy dla liniowości modelu. Aby rozwiązać te problemy, można spróbować zmodyfikować dane, np. za pomocą logarytmowania.