AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Ekonometria

Projekt 6

Temat: Oszacowanie i weryfikacja parametrów wybranej postaci modelu regresji wielorakiej

Autorzy: Laura Cygan, Wiktoria Bąk

Wydział: Wydział Zarządzania

Kierunek: Informatyka i Ekonometria

Przedmiot: Ekonometria

Kraków, 2024

WSTĘP

Celem finalnej części projektu jest oszacowanie parametrów modelu regresji wielorakiej na podstawie wybranych w poprzednich częściach danych, zweryfikowanie poprawności modelu oraz istotności i stabilności jego parametrów. Rozważamy osobne modele dla trzech grup danych – wszystkich powiatów ogółem, a także osobno dla powiatów grodzkich i ziemskich.

Naszą zmienną objaśnianą jest bezrobocie (w %), a zmiennymi objaśniającymi średnie wynagrodzenie, współczynnik feminizacji, współczynnik urbanizacji, liczba ofert pracy i liczba małżeństw przypadających na 10 tys. mieszkańców.

CZĘŚĆ PROJEKTOWA

Dla wszystkich powiatów

W projekcie 5 ustaliłyśmy, że podstawowy model uwzględniający dane ze wszystkich powiatów nie spełnia dwóch założeń twierdzenia Gaussa-Markowa – występowała w nim heteroskedastyczność, a składniki losowe nie miały rozkładu normalnego. Zaproponowałyśmy zlogarytmowanie zmiennych, co też w tej części projektu wykonałyśmy. Sprawdzałyśmy różne kombinacje, jednakże najlepszy model uzyskałyśmy logarytmując zarówno zmienną objaśnianą, jak i wszystkie zmienne objaśniające.

Poprawność modelu

Dla nowego modelu, czynnik inflacji wariancji (VIF) prezentuje się w następujący sposób:

Wynagrodzenie	Ws. feminizacji	Ws. urbanizacji	Oferty pracy	Liczba małżeństw
1,390	3,300	3,244	1,130	1,048

Mimo wprowadzonych zmian, VIF dla poszczególnych zmiennych pozostaje na niskim poziomie (VIF < 10). Oznacza to, że w modelu nie występuje współliniowość.

Aby wybrać najodpowiedniejszy zestaw zmiennych objaśniających po ich zlogarytmowaniu, zastosowaliśmy metodę regresji krokowej w dwóch wariantach: opartą na statystyce F oraz wykorzystującą AIC, czyli kryterium informacyjne Akaike. Oba warianty jako zmienne

objaśniające wybrały wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy oraz liczbę małżeństw, **odrzucając współczynnik urbanizacji**.

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

Homoskedastyczność

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

P-value: 0,058

Wartość nieznacznie większa od $\alpha = 0,05$ oznacza brak podstaw do odrzucenia hipotezy zerowej, czyli w modelu występuje **homoskedastyczność**.

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 Rainbow.

P-value dla testu RESET: 0,341

P-value dla testu Rainbow: 0,031

W przypadku testu RESET p-value znacząco przekracza poziom istotności $\alpha = 0,05$, co sugeruje brak podstaw do odrzucenia hipotezy zerowej o liniowości modelu. Bardziej skomplikowana sprawa jest z wynikiem testu Rainbow, gdyż p-value = 0,03 sugeruje odrzucenie hipotezy zerowej na poziomie istotności $\alpha = 0,05$. Jednakże, przy założeniu innego poziomu istotności, np. $\alpha = 0,01$, wynik testu będzie odwrotny. Na poziomie istotności $\alpha = 0,01$ sprawdzany model **jest więc liniowy.**

Normalność składników losowych

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

P-value: 0.055

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**.

Oszacowanie parametrów i ich istotność

Aby sprawdzić oszacowanie i istotność parametrów, użyliśmy w RStudio komendy summary(model). W poniżej tabeli przestawione są współczynniki przy zmiennych objaśniających oraz wartości p-value przeprowadzonego testu t.

	Stała	Wynagrodzenie	Wsp.feminizacji	Oferty pracy	Małżeństwa
Współczynnik	20,832	-0,695	-2,083	-0,155	-0,912
p-value	2,31e-11	0,002	0,002	0,001	2,27e-05

Jak widać, wszystkie współczynniki zostały uznane za istotne (p-value < 0.05). Można więc uznać, że istnieje istotna statystycznie zależność między wszystkimi zmiennymi objaśniającymi a poziomem bezrobocia.

Za pomocą tej samej komendy w R możemy odczytać również:

- Współczynnik determinacji R^2 = 0,18. Oznacza to, że zmienne niezależne w analizowanym modelu wyjaśniają około 18% zmienności zmiennej zależnej (bezrobocia).
- Skorygowany współczynnik R^2 = 0,17, który wynosi nieznacznie mniej niż zwykły R^2.
- Wartość statystyki F i p-value = 2.226e-14. Test F służy do oceny ogólnej istotności statystycznej całego modelu. Przy tak niskim p-value możemy odrzucić hipotezę zerową, co sugeruje, że co najmniej jedna z analizowanych zmiennych niezależnych jest istotnym predyktorem bezrobocia.

Stabilność parametrów

Aby sprawdzić stabilność parametrów modelu postanowiłyśmy wykorzystać **test Chowa**. Podzieliłyśmy dane na dwie grupy: dane z powiatów grodzkich i dane z powiatów ziemskich.

P-value: 0,12

Wartość większa od $\alpha = 0,05$ oznacza brak podstaw do odrzucenia hipotezy zerowej mówiącej o stabilności parametrów, czyli **parametry modelu są stabilne**.

Dla powiatów ziemskich

W projekcie 5 ustaliłyśmy, że podstawowy model uwzględniający dane z powiatów ziemskich nie spełnia dwóch założeń twierdzenia Gaussa-Markowa – występowała w nim heteroskedastyczność, a składniki losowe nie miały rozkładu normalnego. Zaproponowałyśmy zlogarytmowanie zmiennych, co też w tej części projektu wykonałyśmy. Ponownie, sprawdzałyśmy różne kombinacje, jednakże najlepszy model uzyskałyśmy logarytmując zarówno zmienną objaśnianą, jak i wszystkie zmienne objaśniające.

Poprawność modelu

Dla nowego modelu, czynnik inflacji wariancji (VIF) prezentuje się w następujący sposób:

Wynagrodzenie	Ws. feminizacji	Ws. urbanizacji	Oferty pracy	Liczba małżeństw
1,149	1,661	1,627	1,099	1,163

Mimo wprowadzonych zmian, VIF dla poszczególnych zmiennych pozostaje na niskim poziomie (VIF < 10). Oznacza to, że w modelu nie występuje współliniowość.

Ponownie, aby wybrać najodpowiedniejszy zestaw zmiennych objaśniających po ich zlogarytmowaniu, zastosowaliśmy metodę regresji krokowej w dwóch wariantach: opartą na statystyce F oraz wykorzystującą AIC, czyli kryterium informacyjne Akaike. Oba warianty jako zmienne objaśniające wybrały wynagrodzenie, współczynnik feminizacji, liczbę ofert pracy oraz liczbę małżeństw, **odrzucając współczynnik urbanizacji**.

Mając wybrane odpowiednie zmienne objaśniające do nowego modelu sprawdziliśmy, czy spełniają one założenia twierdzenia Gaussa-Markowa.

Homoskedastyczność

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

P-value: 0.124

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 Rainbow.

P-value dla testu RESET: 0.299

P-value dla testu Rainbow: 0.034

W przypadku testu RESET p-value znacząco przekracza poziom istotności $\alpha = 0{,}05$, co sugeruje brak podstaw do odrzucenia hipotezy zerowej o liniowości modelu. Bardziej skomplikowana sprawa jest z wynikiem testu Rainbow, gdyż p-value = 0,03 sugeruje odrzucenie hipotezy zerowej na poziomie istotności $\alpha = 0{,}05$. Jednakże, przy założeniu innego poziomu istotności, np. $\alpha = 0{,}01$, wynik testu będzie odwrotny. Na poziomie istotności $\alpha = 0{,}01$ sprawdzany model **jest więc liniowy.**

Normalność składników losowych

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

P-value: 0.139

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

Oszacowanie parametrów i ich istotność

Aby sprawdzić oszacowanie i istotność parametrów, użyliśmy w RStudio komendy summary(model). W poniżej tabeli przestawione są współczynniki przy zmiennych objaśniających oraz wartości p-value przeprowadzonego testu t.

	Stała	Wynagrodzenie	Wsp.feminizacji	Oferty pracy	Małżeństwa
Współczynnik	32,801	-0,999	-3,985	-0,125	-1,105
p-value	1,10e-08	0,001	0,001	0,011	1,32e-05

Jak widać, wszystkie współczynniki zostały uznane za istotne (p-value < 0.05). Można więc uznać, że istnieje istotna statystycznie zależność między wszystkimi zmiennymi objaśniającymi a poziomem bezrobocia.

Za pomocą tej samej komendy w R możemy odczytać również:

- Współczynnik determinacji R^2 = 0,16. Oznacza to, że zmienne niezależne w analizowanym modelu wyjaśniają około 16% zmienności zmiennej zależnej (bezrobocia).
- Skorygowany współczynnik R^2 = 0,14, który wynosi nieznacznie mniej niż zwykły R^2.
- Wartość statystyki F i p-value = 7.16e-10. Test F służy do oceny ogólnej istotności statystycznej całego modelu. Przy tak niskim p-value możemy odrzucić hipotezę zerową, co sugeruje, że co najmniej jedna z analizowanych zmiennych niezależnych jest istotnym predyktorem bezrobocia.

Dla powiatów grodzkich

W przypadku danych wyłącznie dla powiatów grodzkich pierwotny model spełniał założenia twierdzenia Gaussa-Markowa. Z tego powodu nie modyfikowaliśmy go, a poniżej ponownie przedstawiona została weryfikacja poprawności modelu z projektu 5.

Poprawność modelu

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 (VIF < 10) i sugerują brak współliniowości.

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 Rainbow.

P-value dla testu RESET: 0.02389493

P-value dla testu Rainbow: 0.05392377

W przypadku testu Rainbow p-value nieznacznie przekracza poziom $\alpha = 0.05$, natomiast dla testu RESET p-value wynosi jedynie 0.02, co sugeruje odrzucenie hipotezy zerowej na poziomie istotności $\alpha = 0,05$. Jednakże, przy założeniu innego poziomu istotności, np. $\alpha = 0,01$, wynik testu będzie odwrotny. Na poziomie istotności $\alpha = 0,01$ sprawdzany model **jest więc liniowy.**

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**.

Oszacowanie parametrów i ich istotność

Aby sprawdzić oszacowanie i istotność parametrów, użyliśmy w RStudio komendy summary(model). W poniżej tabeli przestawione są współczynniki przy zmiennych objaśniających oraz wartości p-value przeprowadzonego testu t.

	Stała	Wynagrodzenie	Wsp. feminizacji	Oferty pracy	Liczba małżeństw
Współczynnik	– 1,3268	– 0,0003	0,0641	– 0,0003	- 0,0341

p-value	0,762	0,048	0,099	0,452	0,147

Jak widać, za istotne zostały uznane jedynie współczynniki przy zmiennej „Wynagrodzenie” (na poziomie istotności 0,05) i zmiennej „Współczynnik feminizacji” (na poziomie istotności 0,1). Można więc uznać, że istnieje istotna statystycznie zależność między wynagrodzeniem a poziomem bezrobocia i współczynnikiem feminizacji a poziomem bezrobocia.

Za pomocą tej samej komendy w R możemy odczytać również:

- Współczynnik determinacji R^2 = 0,23. Oznacza to, że zmienne niezależne w analizowanym modelu wyjaśniają około 23% zmienności zmiennej zależnej (bezrobocia).
- Skorygowany współczynnik R^2 = 0,17, który wynosi nieznacznie mniej niż zwykły R^2.
- Wartość statystyki F i p-value = 0,006. Test F służy do oceny ogólnej istotności statystycznej całego modelu. Przy p-value równym 0,006 możemy odrzucić hipotezę zerową, co sugeruje, że co najmniej jedna z analizowanych zmiennych niezależnych jest istotnym predyktorem bezrobocia.

PODSUMOWANIE

Ostatecznie, zmienne we wszystkich modelach spełniają założenia twierdzenia Gaussa-Markowa. W modelach dla wszystkich powiatów i powiatów ziemskich musieliśmy zlogarytmować zarówno zmienną objaśnianą, jak i wszystkie zmienne objaśniające, a wszystkie parametry uznane zostały za istotne. Niestety, w modelu dla danych z powiatów grodzkich, za istotne zostały uznane jedynie współczynniki przy zmiennych „Wynagrodzenie” i „Współczynnik feminizacji”. Współczynnik determinacji R^2 w każdym przypadku oscyluje na poziomie ok. 15-20%, co nie jest zadowalającym wynikiem. Wszystkie modele są jednak statystycznie istotne (test F).