Tomasz Dyrka
nr indeksu:
studia stacjonarne magisterskie w SGH

przedmiot: Ekonometeria bayesowska
prowadzący: dr Andrzej Torój
data: 28 stycznia 2017r.

Praca zaliczeniowa z przedmiotu

Ekonometria bayesowska

Determinanty stopy bezrobocia w krajach Unii Europejskiej

1 Wstęp

Stopa bezrobocia jest jednym z podstawowym wskaźników ekonomicznych obrazujących ogólnie rozumiany poziom *jakości życia* w danym regionie. W związku z tym może być ona obiektem zainteresowania chociażby rządu, mającego na celu poprawę jakości życia społeczeństwa poprzez zmniejszenie jej wielkości. W takim przedsięwzięciu kluczowe jest zrozumienie natury i przyczyn zjawiska, jakim jest bezrobocie. Możliwe jest zbadanie relacji pomiędzy stopą bezrobocia a innymi wskaźnikami ekonomicznymi, na które rząd ma wpływ w sposób bezpośredni (liczba tworzonych miejsc pracy w firmach państwowych) lub pośredni (dostosowanie przepisów w taki sposób, aby zwiększyć popyt na nowych pracowników na rynku). Takimi wskaźnikami może być na przykład siła związków zawodowych, stopa inflacji, charakter zatrudniania pracowników (umowa o pracę lub o dzieło), płaca minimalna czy ustalanie wysokości zasiłków dla bezrobotnych.

W poniższej pracy zostanie przedstawiony bayesowski model ekonometryczny, który wyjaśni, jak silnie wpływają poszczególne czynniki na wysokość stopy bezrobocia w krajach Unii Europejskiej w 2013 roku. Zostaną wykorzystane ogólnodostępne dane ze stron internetowych Eurostatu i OECD, dotyczące różnych wskaźników makroekonomicznych oraz wyniki analizy przeprowadzonej przez panią Desiree Tercek z John Carroll University pt.: *Determinants of European and United States Unemployment*, które stanowią podstawę analizy a priori (w pracy określenie *artykuł* będzie się zawsze odnosiło właśnie do tego artykułu).

Pracę wykonano w języku R oraz dostępnej w nim bibliotece *knitr*, która umożliwia jednoczesną pracę w językach R oraz LaTeX.

2 Zbiór danych

2.1 Opis słowny danych

Dane niezbędne do przeprowadzenia analizy zostały wybrane na podstawie *artykułu* (wybrane zmienne objaśniające w opisanym w *artykule* modelu w istotny sposób wpłynęły na wartość zmiennej objaśnianej).

W analizie wykorzystano następujące wskaźniki ekonomiczne (w nawiasach znajduje się skrót zmiennej wykorzystywany w skryptach R oraz angielska nazwa zmiennej pochodząca z artykułu):

1. stopa bezrobocie (ur - unemployment rate)[1]

2. liczba pracowników pracujących na mniej niż cały etap (pte - part time employment)[1] (co ciekawe, dane dotyczące zatrudnionych na mniej niż cały etat są dostępne również na stronie OECD i znacznie różnią się od tych prezentowanych przez Eurostat - wybór danych z Eurostatu nie był podparty racjonalnymi przesłankami, ponieważ założono, że dane pochodzące z obu tych stron są tak samo dobrej jakości i nie ma znaczenia, które zostaną wykorzystane)

3. średnia liczba przepracowanych godzin rocznie przez pojedynczego pracownika (aah - average annual hours actually worked per worker)[2]

4. wydatki państwa na zasiłki dla bezrobotnych - jako procent PKB (pue - public unemployment social expenditure)[2]

5. wydatki socjalne państwa - jako procent PKB (pse - public total social expenditure)[2]

[1] http://ec.europa.eu/eurostat
[2] http://www.data.oecd.org

6. wskaźnik ochrony pracowników przed zwolnieniami (sor - strictness of employment protection - individual and collective dismissals)[3]

7. wskaźnik ochrony pracowników zatrudnianych na kontrakty (sot - strictness of employment protection - temporary contracts)[3]

8. gęstość związków zawodowych - odsetek pracowników zrzeszonych w związkach zawodowych (tud - trade union density)[3]

Niestety dane znajdujące się na stronach internetowych OECD oraz Eurostatu nie były kompletne. Z tej przyczyny do analizy wykorzystano dane z 2013 roku (były to najnowsze kompletne dane). Poza tym nie wykorzystano informacji dotyczących Chorwacji, Cypru, Litwy, Łotwy, Malty, Rumunii i Słowacji. Wartości zmiennej *tud - gęstość związków zawodowych* dla Luksemburgu, Polski i Portugalii pochodzą z 2012 roku (założono, że wartość tej zmiennej nie zmieniła się znacznie w latach 2012-2013). Brak wartości zmiennej *pte - part time employment* dla Francji w 2013r. został zastąpiony wartością z 2014 roku.

2.2 Opis statystyczny danych

W tabeli 1 przedstawiono zbiór danych.

	country	ur	pte	aah	pue	pse	sor	sot	tud
1	Belgium	8.4	24.1	1558.0	3.2	29.3	1.9	2.4	55.1
2	Czech Republic	7.0	5.7	1763.0	0.6	20.3	2.9	1.4	12.7
3	Denmark	7.0	20.9	1457.0	0.0	29.0	2.2	1.4	66.8
4	Germany	5.2	26.7	1361.7	1.0	24.8	2.7	1.1	18.1
5	Estonia	8.6	8.7	1866.0	0.3	15.9	1.8	3.0	5.7
6	Ireland	13.1	22.7	1815.0	2.5	20.2	1.4	0.6	29.6
7	Greece	27.5	8.3	2063.0	0.5	26.0	2.1	2.2	21.5
8	Spain	26.1	15.5	1695.6	3.1	26.3	2.0	2.6	16.9
9	France	10.3	18.5	1474.3	1.6	31.5	2.4	3.6	7.7
10	Italy	12.1	17.6	1719.5	1.7	28.6	2.7	2.0	37.3
11	Luxembourg	5.9	18.4	1503.0	1.4	23.2	2.2	3.8	32.8
12	Hungary	10.2	6.4	1879.8	0.5	22.1	1.6	1.2	10.5
13	Netherlands	7.3	46.9	1415.0	1.6	22.9	2.8	0.9	17.8
14	Austria	5.4	26.5	1636.7	1.0	18.1	2.4	1.3	27.8
15	Poland	10.3	6.9	1918.0	0.2	19.6	2.2	1.8	12.7
16	Portugal	16.4	10.8	1859.0	1.6	25.5	3.2	1.8	18.9
17	Slovenia	10.1	8.5	1655.0	0.7	24.0	2.6	1.8	21.2
18	Finland	8.2	12.5	1640.0	1.9	29.5	2.2	1.6	69.0
19	Sweden	8.0	23.4	1609.0	0.5	27.4	2.6	0.8	67.7
20	United Kingdom	7.6	24.2	1666.0	0.3	21.9	1.1	0.4	25.8

Tabela 1: Zbiór danych

Dane dotyczą 20 państw Unii Europejskiej. Zmienną objaśnianą charakteryzują statystyki przedstawione w tabeli 2.

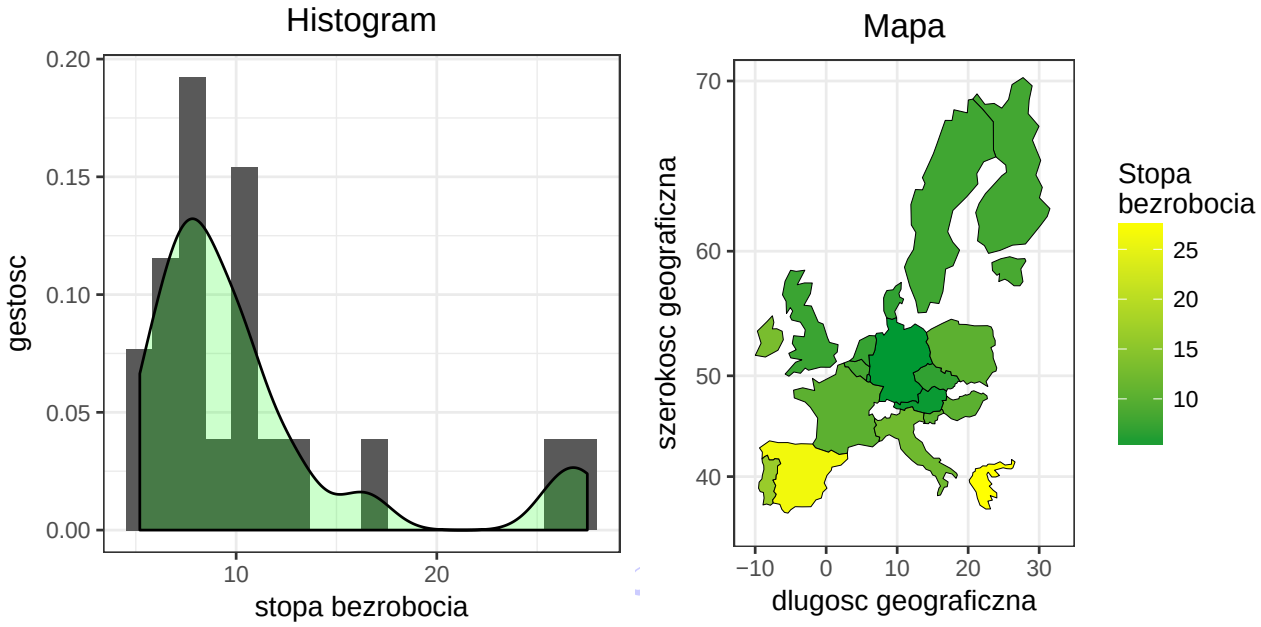
Ze statystyk przedstawionych w tabeli 2 wynika, że średnia stopa bezrobocia w krajach Unii Europejskiej wynosi 10.735 przy - mogłoby się wydawać - dosyć dużym odchyleniu standardowym wynoszącym 6.13. Wstępnie można wysnuć wniosek, że Unia Europejska jest dosyć silnie zróżnicowana pod względem jakości życia.

[3]http://stats.oecd.org

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1.00	20.00	10.73	6.13	8.50	9.41	2.59	5.20	27.50	22.30	1.71	1.92	1.37

Tabela 2: Podstawowe statystyki zmiennej objaśnianej

Dla uplastycznienia zmiennej objaśnianej zostaje ona przedstawiona na rysunku 1.



(a) Histogram oraz przybliżona gęstość

(b) Mapa państw ujętych w analizie

Rysunek 1: Wizualizacja stopy bezrobocia w krajach Unii Europejskiej

Na histogramie widoczne jest pewne podobieństwo do rozkładu normalnego. Zaburza je wysoki poziom bezrobocia w Hipszanii i Grecji (po ponad 20% - jest to widoczne na mapie) oraz brak obserwacji ujemnych (stopa bezrobocia z definicji jest nieujemna, więc rozkład jest ucięty w zerze). Gdyby pominąć te dwa zjawiska oraz gdyby dysponowano większą próbą, histogram najpewniej jednoznacznie by wskazywał, że zmienna jest zbieżna do rozkładu normalnego.

3 Wykorzystany model

Do badania stopy bezrobocia w poszczególnych krajach Unii Europejskiej można wykorzystać model panelowy, który uwzględni informacje z poprzednich kilkunastu lat (lub nawet kwartałów). Dzięki temu można uzyskać *lepsze* oszacowania ze względu na większą próbę[4]. Mankamentem takiego podejścia jest odrzucenie wyników badań, które zostały już przeprowadzone na temat determinant stopy bezrobocia.

Charakterystyczna dla modelowania bayesowskiego jest możliwość połączenia wiedzy eksperckiej z dostępnymi danymi w jeden spójny model. Wiedza ekspercka, określana jako *a priori*,

[4]Próbowałem stworzyć bayesowski model panelowy na tych samych danych (oczywiście z dodatkowym wymiarem czasu), który mógłbym oszacować za pomocą funkcji HMMpanelFE z pakietu MCMCpack, jednak korzystając z funkcji BayesFactor w tym pakiecie nie udało mi się uzyskać wartości czynnika Bayesa (wyskakiwał error). Stając przed dylematem: analiza danych panelowych bez czynnika Bayesa lub danych przekrojowych z czynnikiem, wybrałem drugą opcję.

jest to pewne domniemanie dotyczące rokładów parametrów modelu (w szczególności ich wartości oczekiwanych, wariancji i korelacji między nimi), które wynika z doświadczenia badacza lub innych przesłanek, których nie sposób potraktować jako dane ilościowe. Z kolei dostępne dane są interpretowane w następujący sposób: jakie jest prawdopodobieństwo wylosowania właśnie takich wartości, przy różnych wartościach parametrów. Taka probabilistyczna interpretacja danych nazywana jest *gęstością danych*. Połączenie gęstości danych i informacji a priori pozwala na uzyskanie oczasowań parametrów *a posteriori*, co przy wykorzystaniu twierdzenia Bayesa można zapisać w następujący sposób:

$$f(\theta|y) \propto f(y|\theta) \cdot f(\theta) \tag{1}$$

Warto zwrócić uwagę, że nie jest to równość, jednak pominięcie elementu skalującego nie wpłynie na interpretację parametrów a posteriori.

3.1 A priori

W modelu wykorzystanym w poniższej analizie założono a priori (zgodnie z *artykułem*), że stopa bezrobocia jest liniową kombinacją zmiennych objaśniających. W związku z tym postać funkcyjna modelu przedstawia się następująco:

$$ur_i = \beta_0 + \beta_1 pte_{1,i} + \beta_2 aah_{2,i} + \beta_3 pue_{3,i} + \beta_4 pse_{4,i} + \beta_5 sor_{5,i} + \beta_6 sot_{6,i} + \beta_7 tud_{7,i} + \varepsilon_i \tag{2}$$

Ponadto przyjęto, że składnik losowy oraz parametry mają rozkłady normalne o wartościach oczekiwanych równych odpowiednio: 0 i oszacowaniom parametrów w *artykule* oraz wariancjach σ^2 i błędom standardowym oszacowań parametrów podniesionym do kwadratu (dane z *artykułu*). Aby opisać rozkład wariancji składnika losowego, posułożono się zmienną pomocniczą h, którą można w jednoznaczny sposób przekształcić w σ^2 ($h^{-1} = \sigma^2$). Założono, że ma ona rozkład gamma. Powyższe informacje można zapisać w następującej formie (uwzględniono, że h występuje w rozkładzie bet jako parametr, co oznacza, że rozkład bet jest zależny od wartości h):

$$\boldsymbol{\beta}|h \sim \boldsymbol{N}(\underline{\boldsymbol{\beta}}, h^{-1}\underline{\boldsymbol{U}}) \tag{3}$$

$$h \sim G(\underline{s}^{-2}, \underline{\nu}) \tag{4}$$

co można zapisać jako jeden rozkład normalny-gamma:

$$\boldsymbol{\beta}, h \sim \boldsymbol{N}G(\boldsymbol{\beta}, h^{-1}\underline{\boldsymbol{U}}, \underline{s}^{-2}, \underline{\nu}) \tag{5}$$

gdzie $\underline{\boldsymbol{\beta}} = [0, 0, 0.015, 4.201, 0.603, -3.623, -2.981, 0.117]^T$, natomiast
$diag(\underline{h^{-1}\boldsymbol{U}}) = [500, 10, 0.000003, 2.374946, 0.015331, 0.584965, 0.299619, 0.000705]$.

Wariancje stałej i zmiennej *pte* przyjęto jako nieznane a priori, ponieważ w *artykule* stała w modelu była nieistotna statystycznie, natomiast zarówno oszacowanie, jak i wariancja zmiennej *pte* były bardzo bliskie zera (w artykule zmienna *pte* była wyrażona w liczbie pracowników, a nie w postaci odsetka osób pracujących nie na cały etat wśród wszystkich pracujących. Wydaje się to być niezbyt dobrym pomysłem, ponieważ kraje różnią się liczbą ludności, a zmienna objaśniana jest wyrażona jako odsetek.) Wariancja a priori może wydawać się bardzo niska - wynika to z faktu, że same oszacowania paramatrów są małymi (bliskimi zera) wartościami, a ich błędy (odchylenia) standardowe są mniejsze od 1 - te podniesione do kwadratu (w celu otrzymania wariancji) dają bardzo małe wartości. Ponadto, z uwagi na brak informacji w *artykule* na temat sumy kwadratów reszt i, co za tym idzie, wariancji składnika losowego, zostaje ona przyjęta jako 1 (arbitralnie, z uwagi na łatwość obliczeń). W rezultacie parametry rozkładu

a prio wariancji składnika losowego wynoszą: $\underline{s}^{-2} = 1$ oraz $\underline{v} = 92$ (wielkość próby w badaniu w *artykule*). Przy tym wartość oczekiwana rozkładu gamma wynosi $s^{-2} = 1$, więc przyjęto, że we wzorze na wariancje a priori parametrów beta $\underline{\boldsymbol{U}} = h^{-1}\underline{\boldsymbol{U}}$ (umożliwi to wykorzystanie funkcji MCMCregress z pakietu MCMC, która zakłada niezależność rozkładów bet i σ^2, przez co wariancja bet jest zdefiniowana po prostu jako $\underline{\boldsymbol{U}}$, a nie jako $h^{-1}\underline{\boldsymbol{U}}$).

Elementy macierzy $h^{-1}\underline{\boldsymbol{U}}$ leżące poza przekątną (kowariancje parametrów) zostały przyjęte a priori jako zerowe, gdyż nie ma podstaw, by sądzić inaczej (w *artykule* nie znalazła się informacja o kowariancjach oszacowań parametrów).

3.2 Gęstość danych

Obliczenie gęstości danych jest konieczne, aby uzyskać rozkład a posteriori parametrów. Sprowadza się ono do obliczenia funkcji wiarygodności zbioru danych - założono, że wartości stopy bezrobocia są niezależne pomiędzy poszczególnymi krajami i wynikają jedynie z wartości zmiennych objaśniających oraz składnika losowego (który z założenia jest IID). To założenie może budzić kontrowersje z uwagi na fakt, że na terenie Unii Europejskiej siła robocza może przemieszczać się bez ograniczeń, a więc zmniejszenie się bezrobocia w jednym kraju spowodowane emigracją bezrobotnych (na przykład z Polski) może spowodować większe bezrobocie w Wielkiej Brytanii (brytyjscy pracodawcy chętniej zatrudniają tanich pracowników z Polski niż Brytyjczyków, którzy przez to trafiają na bezrobocie). Co więcej, na rysunku 1b widoczne jest, że bardzo wysokie bezrobocie dotyka krajów leżących na południu (Hiszpania, Grecja), co może być związane z klimatem. Być może odpowiednie byłoby zastosowanie tutaj modelu uwzględniającego te wpływy, na przykład ekonometrycznego modelu przestrzennego.

Zgodnie z założeniem dotyczącym wzajemnej niezależności od siebie poszczególnych obserwacji zmiennej objaśnianej, gęstość danych można zapisać jako iloczyn funkcji gęstości każdej obserwacji:

$$f(ur|\boldsymbol{\beta}, h) = \prod_{i=1}^{20} f(ur_i|\boldsymbol{\beta}, h) = \frac{h^{10}}{(2\pi)^{10}} exp[-\frac{h}{2}\sum_{i=1}^{20}(ur_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2] \tag{6}$$

3.3 A posteriori

Obliczenie wartości a posteriori jest w przypadku przedstawionego modelu możliwe do wykonania analitycznie, stosując wzór 1. Do obliczenia rozkładów a posteriori wykorzystano następujące wzory[5]

$$\overline{\boldsymbol{\beta}} = (\underline{\boldsymbol{U}}^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1}(\underline{\boldsymbol{U}}^{-1}\underline{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{7}$$

$$\overline{\boldsymbol{U}} = (\underline{\boldsymbol{U}}^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{8}$$

$$\overline{\nu} = \underline{\nu} + N \tag{9}$$

$$\overline{\nu s}^2 = \hat{\nu}\hat{s}^2 + \underline{\nu s}^2 + (\hat{\boldsymbol{\beta}} - \underline{\boldsymbol{\beta}})^T[\underline{\boldsymbol{U}} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}]^{-1}(\hat{\boldsymbol{\beta}} - \underline{\boldsymbol{\beta}}) \tag{10}$$

Zauważono, że rozkład a posteriori jest również rozkładem normalnym-gamma, co oznacza, że jest on sprzężony do rozkładu a priori. Aby móc intuicyjnie zinterpretować uzyskane rozkłady a posteriori, należy uniezależnić rozkłady bet od rozkładu h poprzez całkowanie uzyskanego rozkładu a posteriori po h (innymi słowy oznacza to policzenie gęstości brzegowej rozkładu). W rezultacie z początkowego rozkładu bet (wielowymiarowego rozkładu normalnego) powstaje wielowymiarowy rozkład t-Studenta.

Każdy z rozkładów bet a posteriori ma rozkład t-Studenta, który można zapisać jako:

$$\beta_i|y \sim t(\overline{\beta}_i, \overline{s}^2\overline{U}_{i,i}, \overline{v}_i) \tag{11}$$

[5]wzory pochodzą z wykładu 4 *Bayesowska analiza modelu regresji liniowej wielu zmiennych*

Na rysunku 2 przedstawiono rozkłady a posteriori bet (niezależne od rozkładu h) zestawione z rozkładami a priori.

W modelach bayesowskich równanie opisujące zmienną objaśnianą nie zawsze jest liniowe, a parametry nie zawsze są zbieżne do typowego rozkładu (jakimi są na przykład rozkład normalny lub gamma). W takich sytuacjach jednym ze sposobów uzyskania rozkładu a posteriori jest zastosowanie metody numerycznej klasy MCMC (Markov Chain Monte Carlo), które umożliwiają n-krotne losowanie z rozkładu a posteriori, mając dane rozkłady a priori i gęstość danych. Jednym z takich algorytmów jest próbnik Gibbsa. W R dostępna jest funkcja MCMCregress z bibliotece MCMCpack, która umożliwia zastosowanie próbnika Gibbsa do przybliżenia rozkładu a posteriori parametrów modelu liniowego o analogicznej specyfikacji do przedstawionego powyżej, z jednym wyjątkiem: założono a priori, że rozkłady bet oraz wariancji składnika losowego są od siebie niezależne. W rezultacie rozkłady bet a posteriori będą miały po prostu rozkłady normalne - nie będzie konieczne liczenie gęstości brzegowej. Pomimo to na wykresie 2 przedstawiono rozkłady parametrów a posteriori szacowane za pomocą MCMCregress, aby zweryfikować, na ile założenie o zależności bet od σ^2 ma wpływa na rozkład a posteriori oraz na ile skuteczny jest próbnik Gibbsa - można to sprawdzic porównując wartości MCMC z obliczonymi analitycznie.

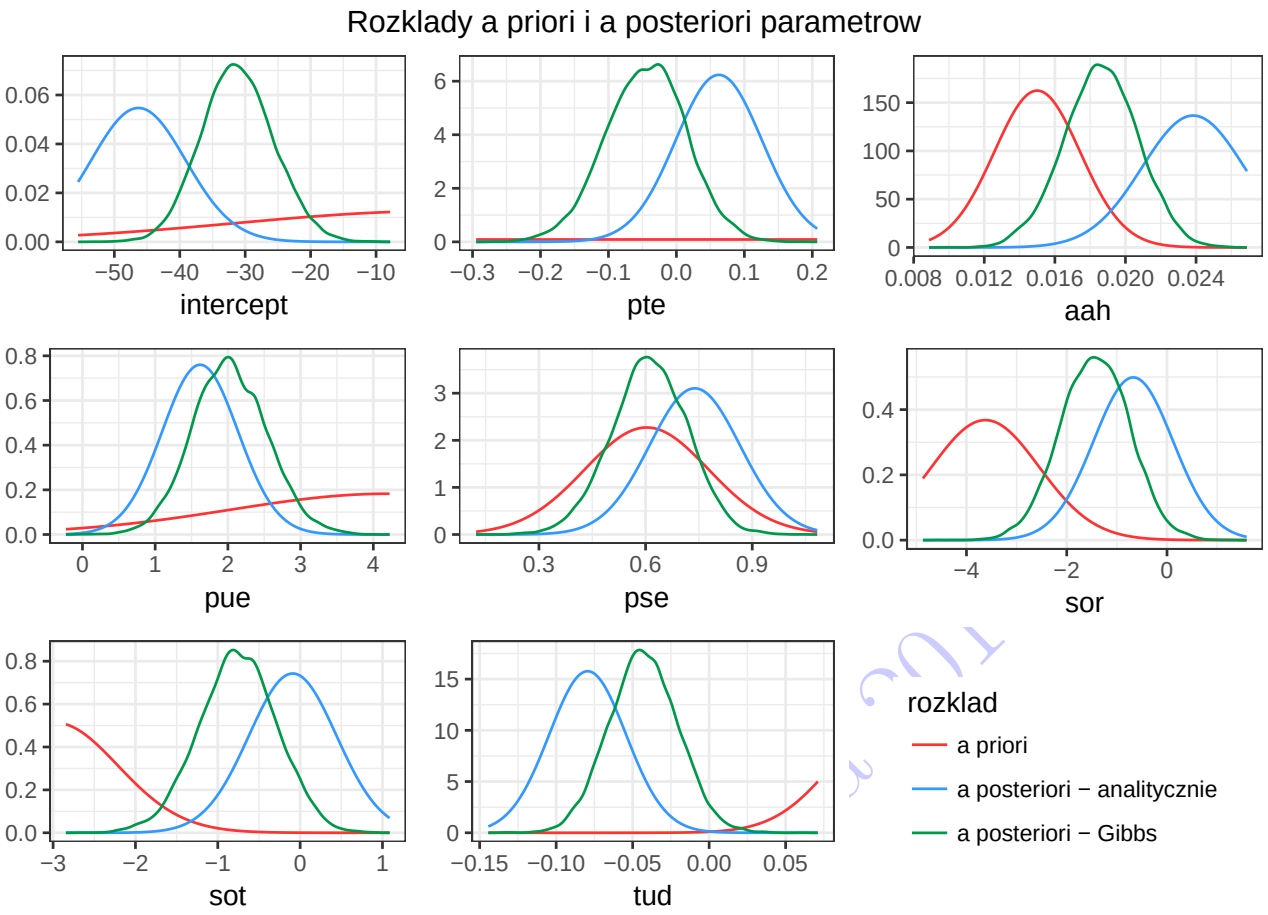
4 Ocena jakości modelu

4.1 Porównanie a priori z a posteriori

Celem modelowania bayesowskiego jest oczywiście uzyskanie modelu, który *dobrze* opisuje badaną zmienną, jednak właściwie niemożliwe jest znalezienie obiektywnej miary, która odpowie, jak *dobry* jest model. W takiej sytuacji za sukces analityka można uznać sytuację, kiedy w wyniku analizy parametry a posteriori maja mniejszą wariancję niz a priori.

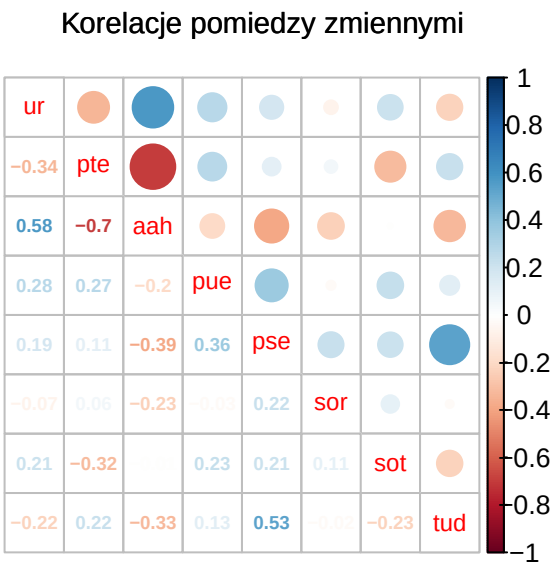
W tym miejscu wprowadzono poprawkę do rozkładów a priori, co jest niezgodne ze sztuką tworzenia modelu bayesowskiego. Wynikało to z uzyskania modelu, w którym w 5 na 8 zmiennych wariancja a posteriori była większa niż a priori, czego przyczyną mogło być przyjęcie zbyt małych wariancji a priori. Z kolei przyczyną dla przyjęcia tak małych wartości była intuicja autora pracy, która najwidoczniej - z uwagi na małe doświadczenie w tworzeniu modelu bayesowskich - zawiodła. Nowa wariancja a priori jest dwukrotnie wyższa od wcześniej przyjetej i od tej chwili analiza będzie oparta o nowe wartości wariancji.

Na rysunku 2 widoczne jest, że dla wszystkich zmiennych (z wyjątkiem aah) doszło do zawężenia wariancji po wprowadzeniu do modelu danych, co można uznać za sukces. Nie jest to zaskakujące dla zmiennych intercept oraz pte, gdzie a priori celowo przyjęto wysoką wartość wariancji (w ten sposób niemal całość informacji dotyczącej rozkładu tych parametrów a posteriori pochodzi z danych). Widoczne jest również podobieństwo rozkładu a priori do a posteriori parametru zmiennej pse (wydatki państwa na opiekę społeczną jako % PKB). Oznacza to, że rozkład a priori był podobny do tego zawartego w danych. Dla zmiennych pue (zasiłki dla bezrobotnych jako % PKB), sor (ochrona pracowników stałych), sot (ochrona pracowników na kontrakty) oraz tud (gęstość związków zawodowych) widoczne jest znaczne zmniejszenie wariancji. Zmienne sor oraz sot (ochrona pracowników stałych i na kontraktach) przestały w znaczący sposób wpływać na stopę bezrobocia, ponieważ wartości oczekiwane ich parametrów znacznie zbliżyły się do zera.



Rysunek 2: Porównanie rozkładów a priori oraz a posteriori uzyskanego analitycznie i za pomocą próbnika Gibbsa

Ciekawa jest również zmiana kierunku wpływu zmiennej tud (gęstość związków zawodowych) z dodatniego na ujemny. Model wskazuje, że im więcej pracowników zrzeszonych jest w związki zawodowe, tym niższe jest bezrobocie (relacja przyczynowo - skutkowa pomiędzy zmiennymi tud i ur (stopa bezrobocia) nie jest oczywista: z jednej strony wysokie bezrobocie może wiązać się z dużym ryzykiem utraty pracy, co zachęca do uczestnictwa w związkach, z drugiej - wysoki odsetek pracowników w związkach utrudnia zwolnienia).



Rysunek 3: Korelacje pomiędzy zmiennymi

Nietypowe zachowanie zmiennej aah (zwiększenie wariancji a posteriori w porównaniu do a priori) może wynikać z występowania silnych korelacji pomiędzy zmiennymi. Na rysunku 3 widoczna jest dosyć silna ujemna korelacja pomiędzy zmiennymi pte (odsetek zatrudnionych nie na cały etat) oraz aah (średnia liczba przepracowanych godzin rocznie przez pracownika), co jest zresztą dosyć oczywiste, gdyż osoby pracujące na mniej niż cały etat pracują krócej. Przy tak silnej korelacji można zastanowić się nad wykluczeniem którejś ze zmiennych z modelu lub policzyć dodatkowy wskaźnik, który mógłby uzasadnić taką decyzję, np. VIF (czynnik inflacji wariancji).

Wyniki uzyskane za pomocą próbnika Gibbsa są dla większości zmiennych podobne do rozwiązania analitycznego, jednak za każdym razem ich wartość oczekiwana jest bliższa wartości oczekiwanej a priori. Dla większości parametrów próbnik Gibbsa wypadł zadowalająco, jednak dla zmiennych sor i sot otrzymane wartości oczekiwane są bardzo bliskie zera, co wskazuje na niemal niezauważalny wpływ tych zmiennych na stopę bezrobocia i może sugerować usunięcie tych zmiennych z modelu.

4.2 Czynnik Bayesa

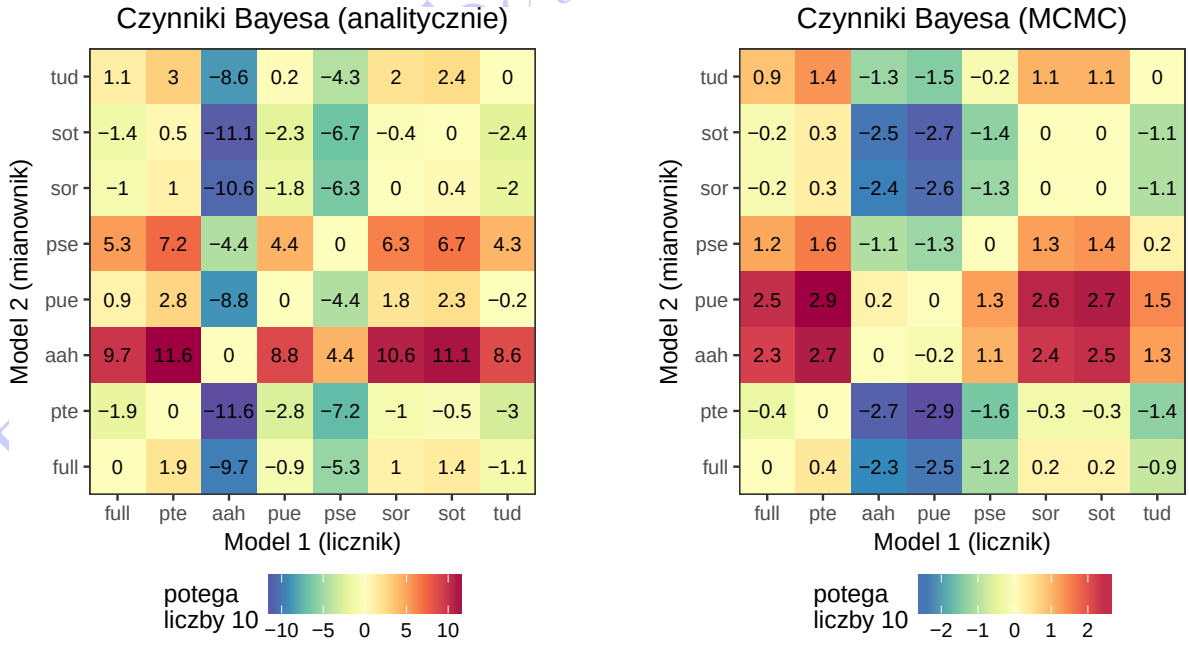
Czynnik Bayesa jest miarą, która pozwala na porównanie ze sobą dwóch modeli i ocenę, który z nich jest *lepszy*. Definiuje się go jako:

$$BF^{(i,j)} = \frac{P(\boldsymbol{y}|M^{(i)})}{P(\boldsymbol{y}|M^{(j)})} = PO^{(i,j)} \frac{P(M^{(i)})}{P(M^{(j)})} \tag{12}$$

czyli jest to iloraz szans a posteriori (szansa na otrzymanie właśnie takich rozkładów parametrów przy określonym $\boldsymbol{y}$) pomnożony przez iloraz szans a priori (czyli ocenę a priori, który z modeli jest *lepszy*). Warto zwrócić uwagę, że w przypadku przyjęcia równych wartości szans a priori dla obu modeli (tzw. nieinformacyjnego a priori), czynnik Bayesa jest równy ilorazowi a posteriori.

Czynnik Bayesa może być stosowany do porównywania dowolnych dwóch modeli wyjaśniających tą samą zmienną. W szczególności jest on przydatny w sytuacji, kiedy porównujemy dwa modele, z których jeden zawiera pewien zestaw zmiennych, a drugi ten sam zestaw pomniejszony o jedną zmienną. Uzyskana wtedy miara może być interpretowana analogicznie do statystyki t-Studenta, badającej istotność zmiennej w modelu w analizie klasycznej.

W analizowanym zagadnieniu czynnik Bayesa może zostać wykorzystany w celu zweryfikowania przydatności w modelu poszczególnych zmiennych: jeżeli okaże się, że model bez danej zmiennej jest *lepszy*, należy usunąć taka zmienną. Przyjęto nieinformacyjne a priori.



(a) Czynniki Bayesa policzone analitycznie (b) Czynniki Bayesa policzone za pomocą MCMC

Rysunek 4: Czynniki Bayesa

W bardziej skomplikowanych niż prezentowany w poniższej pracy modelach często konieczne jest numeryczne obliczenie licznika i mianownika wzoru na czynnik Bayesa, czyli funkcji

wiarygodności brzegowych obu modeli. Można w tym celu wykorzystać metody klasy MCMC, podobnie jak w przypadku przybliżania gęstości a posteriori. Na rysunku 4, obok czynnikóœ policzonych analitycznie, przedstawiono czynniki obliczone za pomocą funkcji BayesFactor z biblioteki MCMCpack.

Obliczone czynniki Bayesa przedstawiono na rysunku 4. Wykresy należy czytać następująco: w kolumnach znajdują się gęstości zmiennych z licznika wzoru na czynnik Bayesa, natomiast w wierszach - z mianownika. Przykładowo, na wykresie 4a pole na przecięciu wiersza aah oraz kolumny full ma kolor ciemnoczerwony, któremu odpowiada wartość 9.7. Oznacza to, że czynnik Bayesa, gdzie M1 (licznik) to model *full*, czyli model zawierający wszystkie zmienne, natomiast M2 (mianownik) to model *pte*, czyli zawierający wszystkie zmienne z wyjątkiem pte, wynosi 10^8. Kolorami na wykresie zostały oznaczone wartości czynników Bayesa, gdzie wartości kolorów odpowiadają potęgom liczby 10. Taka ilustracja wynika z interpretacji czynnika Bayesa, która według skali Jeffreysa[6] zależy od potęgi liczby 10. Przedstawiono ją w tabeli 3.

	BF	potega	interpretacja
1	$< 10^0$	< 0	negative (supports M_2)
2	$10^0 - 10^5$	$0 - 0.5$	barely worth mentioning
3	$10^{0.5} - 10^1$	$0.5 - 1$	substantial
4	$10^1 - 10^{1.5}$	$1 - 1.5$	strong
5	$10^{1.5} - 10^2$	$1.5 - 2$	very strong
6	$> 10^2$	> 2	decisive

Tabela 3: Interpretacja wartości czynnika Bayesa według Jeffreysa

Na podstawie rysunku 4 oraz tabeli 3 można wyciągnąć następujące wnioski (spostrzeżenia):

1. Każdy z wykresów przedstawiających czynniki Bayesa jest antysymetryczny (chociaż oś symetrii biegnie po przeciwległej przekątnej niż w zwykłych macierzach). Jest to ciekawa własność skali Jeffreysa.

2. Oba sposoby liczenia czynnika Bayesa wskazały, że model full (ze wszystkimi zmiennymi) jest zdecydowanie lepszy od modelu bez zmiennej aah (średnia liczba przepracowanych godzin) - jest to o tyle ciekawe, że jedynie ta zmienna miała większą wariancję a posteriori niż a priori, a więc początkowo mogła wydawać się zbędna, wręcz - szkodliwa).

3. Równie *istotną* zmienną jest pse (wydatki soacjalne państwa), chociaż w analizie MCMC model z tą zmienną jest *tylko bardzo silnie* lepszy od modelu bez tej zmiennej, natomiast w analizie analitycznej jest zdecydowanie lepszy.

4. Wątpliwości budzi rozbieżność wyników dotyczących zmiennej pue (wydatki państwa na zasiłki dla bezrobotnych) pomiędzy analizą anlityczną i MCMC. Szczególnie, że parametr a posteriori przy tej zmiennej znaczenie zmniejszył swoją wariancję w porównaniu do a priori w obu analizach.

5. Obie analizy dały podobne wyniki, chociaż analiza MCMC znacznie ostrożniej szacuje wartości czynnika Bayesa.

[6]na podstawie wykładu 5: *Elementy wnioskowania w modelach bayesowskich: HPDI, iloraz szans a posteriori, czynnik Bayesa*