**Bayesowski Model Normalnej Regresji Liniowej – ćwiczenie: komentarz do wyników**

1. Przypomnijmy równanie leżące u podstaw modelu próbkowego:

, ,

gdzie:

* – kwartalna wartość wypłat gotówki w bankomatach w Polsce (w mln zł),
* – zmienna czasowa (), której kolejne wartości oznaczają numery kolejnych obserwacji,
* reprezentują zero-jedynkowe zmienne sezonowe odpowiadające kwartałom II , III i IV ,
* – liczba bankomatów w Polsce (w setkach szt.),
* – liczba wydanych kart płatniczych (w tys. szt.).
1. Liczba obserwacji:

Liczba parametrów strukturalnych (): (łącznie z wyrazem wolnym)

Liczba wszystkich parametrów modelu (łącznie z parametrem zakłócającym; ang. *nuisance parameter*, którym jest ):

1. Przypomnijmy założenia odnośnie rozkładu *a priori*(gamma-normalnego z zależnością pomiędzy i ):
	* i , co oznacza, że (dla każdego ) i , , oraz ,
2. Wykaz charakterystyk brzegowych rozkładów *a posteriori*, które należało uzyskać i porównać na tle ich odpowiedników w rozkładach *a priori*:
* wartości oczekiwane, mediany, modalne (dla danego parametru , wszystkie są sobie równe)
* odchylenia standardowe,
* przedziały najwyższej gęstości *a posteriori* (dla poziomu prawdopodobieństwa 0,95)
* wykres funkcji gęstości rozkładów *a posteriori* i *a priori* (na tym samym wykresie) dla parametru odzwierciedlającego związek pomiędzy liczbą bankomatów w wartością wypłat gotówkowych (i tylko dla niego, żeby nie trzeba było powielać tych samych operacji dla pozostałych współczynników regresji)
1. Analiza wyników:
	1. Analizę zaczynamy od wykresu – to pierwsza rzecz, na którą powinniśmy zwrócić uwagę, gdyż pozwala sobie wyrobić opinię odnośnie:
		1. Regularności kształtu rozkładu *a posteriori* – najbardziej regularny kształt charakteryzuje się jednomodalnością i symetrią (jak ma to miejsce w rozkładzie normalnym czy t-Studenta). Im bardziej nieregularny kształt funkcji gęstości rozkładu *a posteriori* (skośność, wiele modalnych, garby), tym powodować to powinno w nas poczucie większej niepewności wnioskowania (choćby z tego tytułu, że wtedy „rozjeżdżają” się – mniej lub bardziej – trzy główne miary położenia rozkładu: wartość oczekiwana, modalna i mediana, i już nie są sobie równe). W przypadku silnych nieregularności rozkładu, np. gdy rozkład jest „silnie” dwumodalny, wówczas wartość oczekiwana – wtenczas ulokowana gdzieś pomiędzy modalnymi – zupełnie nie odzwierciedla tendencji centralnej rozkładu, bo ta jest rozdzielona na dwie osobne części – związane z tymi dwiema modalnymi)
		2. Podobieństwa i różnic pomiędzy rozkładem *a posteriori* (czyli tym odzwierciedlającym naszą końcową – tj. po wglądzie w dane – wiedzę/obraz niepewności co do danego parametru) a rozkładem *a priori* (czyli tym odzwierciedlającym naszą wstępną – tj. przed wglądem w dane – wiedzę/obraz niepewności co do danego parametru). Im większa różnica pomiędzy jednym a drugim, tym mocniej dane zmodyfikowały naszą wstępną wiedzę / obraz niepewności o parametrze, co zasadniczo należy odbierać pozytywnie, gdyż oznacza to, że dane faktycznie wniosły coś nowego / nową informację o parametrze. Podobieństwo i różnice pomiędzy rozkładem *a posteriori* a rozkładem *a priori* należy oceniać przede wszystkim pod kątem ich wzajemnego:
			* położenia: jeżeli centralne części obydwu rozkładów znajdują się „daleko” od siebie (cokolwiek by to znaczyło – to kwestia uznaniowa, oceniana po prostu „na oko”, spoglądając na wykres), wówczas powiemy, że dane znacząco / „mocno” przesunęły rozkład *a priori*, co zasadniczo oceniamy pozytywnie, gdyż – jak wyjaśniono powyżej – oznacza to, że dane wniosły do modelu/wnioskowania jakąś nową znaczącą informację; zauważmy też, że ta „odległość” pomiędzy centralnymi częściami rozkładów *a posteriori* i *a priori* znajduje swoje odzwierciedlenie / przekłada się na różnice pomiędzy miarami położenia tych rozkładów (wartości oczekiwanych, modalnych i median);
			* rozproszenia: zwykle spodziewamy się, że rozkład *a posteriori* powinien charakteryzować się mniejszym rozproszeniem w porównaniu do rozkładu *a priori*, a to dlatego, że dane (informacja w nich zawarta) powinny się przyczyniać do zmniejszenia *niepewności* naszego wnioskowania o parametrach (im bardziej rozproszony rozkład czy to *a priori*, czy *a posteriori*, tym większą *niepewnością*, jest obarczone wnioskowanie o parametrze na podstawie danego rozkładu, odpowiednio, *przed* czy *po* wglądzie w dane). Czasami jednak – i wbrew powyższej intuicji – może się zdarzyć się, że otrzymany rozkład *a posteriori* jest bardziej rozproszony (tj. ma wyższe rozproszenie) aniżeli rozkład *a priori*. Przyczyną może być… silne skorelowanie dwóch zmiennych objaśniających, czyli problem bliskiej (bo nie dokładnej) ich współliniowości (inaczej: liniowej współzależności). Jeśli bowiem dwie zmienne objaśniające są ze sobą silnie skorelowane (czy to dodatnio, czy ujemnie), wówczas niosą one ze sobą „prawie tę samą informację”, za pomocą której usiłujemy wyjaśnić kształtowanie się zmiennej objaśnianej. Przez to model „może mieć problem” z jednoznacznym wnioskowaniem o parametrach stojących przy tym dwóch (prawie) współliniowych zmiennych, bo ich wpływ na y-ka jest bardzo, bardzo podobny – i to właśnie znajduje swoje odzwierciedlenie w wyższym (od *a priori*) rozproszeniu rozkładu *a posteriori* (powtórzmy: ta duża niepewność wnioskowania o jednym czy drugim parametrze wynika z tego, że modelowi ciężko rozróżnić między sobą wpływ jednej i drugiej zmiennej objaśniającej na zmienną objaśnianą).

🡪 W ćwiczeniu uzyskaliśmy taki wykres:



🡪 Co zauważamy? Patrzymy na różnice pod względem położenie i rozproszenia obydwu rozkładów. W tym pierwszym aspekcie zauważamy, że rozkład *a posteriori* znajduje się daleko na prawo od rozkładu *a priori*, a zatem „dane przesunęły naszą wiedzę/obraz niepewności na prawo” (spójrzmy, że rozkład *a posteriori* jest tak bardzo przesunięty, że prawie w całym zakresie wartości prezentowanych na osi OX, wartość gęstości *a priori* jest praktycznie zerowa – oczywiście zera nie sięga, ale jest już bardzo, bardzo jemu bliska). Możemy zatem stwierdzić, że dane wniosły znacząco różną od *a priori* informację do modelu o parametrze (a to dobrze, bo po to w sumie wykorzystujemy dane w statystyce: żeby wnioskować na podstawie informacji zawartej *w nich*, a nie na podstawie naszego *a priori* wyobrażeniach/założeniach o parametrach). Tę różnicę w położeniu centralnych części obu analizowanych rozkładów potwierdzają także ich charakterystyki położenia, bowiem (a swoją drogą, skąd to wiemy?), podczas gdy (patrz arkusz Excela, jak i tabelka poniżej).

A co odnośnie rozproszenia obydwu rozkładów? Ich porównanie na podstawie tylko wykresu jest tu utrudnione, ponieważ nie widać na nim centralnej części rozkładu *a priori* (nie zmieściła się). Możemy jednak spojrzeć na odpowiednie charakterystyki liczbowe tych rozkładów, a mianowicie na odchylenia standardowe (patrz arkusz Excela, jak i tabelka poniżej): , podczas gdy . Stwierdzamy – ku naszemu zaskoczeniu – że rozproszenie rozkładu *a posteriori* (a tym samym niepewność wnioskowania o parametrze *po* wglądzie w dane) jest większe niż *a priori*. Sytuacja ta, jakkolwiek dość nietypowa (z przyczyn opisanych wcześniej) ma tu jednak swoją konkretną przyczynę – dokładnie tę opisaną powyżej. Otóż zmienna objaśniająca stojąca przy parametrze (czyli = liczba bankomatów) okazuje się silnie (akurat dodatnio) liniowo skorelowana z inną zmienną objaśniającą, a mianowicie = liczbą wydanych kart płatniczych, co łatwo zauważyć na poniższych wykresach (wykres liniowy – po lewej stronie; wykres rozrzutu – po prawej stronie).



Ponadto, fakt silnej liniowej korelacji obydwu zmiennych potwierdza także wartość współczynnika korelacji (obliczona za pomocą funkcji =WSP.KORELACJI(…,…)) wynosząca aż ok. 0,955. Możemy zatem stwierdzić, że owszem, rozproszenie rozkładu *a posteriori* – wbrew intuicji – jest większe od tego założonego *a priori* (jakoby dane zamiast zmniejszyć, to zwiększyły niepewność wnioskowania), lecz wynika to z bardzo silnej liniowej korelacji (czyli prawie liniowej współzależności) zmiennej ze zmienną , wobec czego model ma trudność (objawiającą się dużym rozproszeniem *a posteriori*) z bardziej jednoznacznym wnioskowaniem o wpływie liczby bankomatów na wysokość wypłat gotówkowych z bankomatów.

🡪 Uwaga: w praktyce należałoby się przyjrzeć wykresom skonstruowanym dla *każdego*, a nie tylko wybranego parametru strukturalnego; w ćwiczeniu ograniczyliśmy się tylko do jednego.

* 1. Przechodzimy do analizy charakterystyk liczbowych – w tym celu przyglądamy się tabeli:



Z powyższych charakterystyk liczbowych interesują nas:

* Wartości oczekiwane, modalne i mediany *a posteriori* (dla danego parametru równe sobie, ponieważ…?) – porównujemy je z ich odpowiednikami *a priori*, czyli – także równymi sobie (dla danego ). Ponadto, ponieważ przyjęto (jak to zwykle w modelach regresji), że , zachodzi dla każdego . Z tabelki wynika, że wartości owych charakterystyk w rozkładach *a posteriori* są mniej lub bardziej przesunięte od ich wartości (równej 0) w rozkładach *a priori*, co oznacza, że dane zmodyfikowały (mniej lub bardziej znacząco) nasze wstępne założenia / wstępną wiedzę / wstępny (aprioryczny) obraz niepewności odnośnie położenia poszczególnych parametrów.

Ponadto, przypomnijmy, że (przy pewnych dodatkowych założeniach) każdą z rozważanych tu wielkości (równych sobie): wartość oczekiwaną/modalną/medianę *a posteriori* danego parametru strukturalnego, można traktować jako *optymalną bayesowską ocenę punktową parametru* – i interpretować ją tak samo, jak robi się to z oceną MNK () na gruncie ekonometrii niebayesowskiej („klasycznej”) – to proszę sobie koniecznie przypomnieć!

* Odchylenia standardowe *a posteriori* poszczególnych parametrów strukturalnych, czyli – porównujemy je z odchyleniami standardowymi *a priori* parametrów, czyli , które akurat wynoszą 10 dla każdego (ponieważ w rozkładzie *a priori* przyjęto ). Widzimy, że uzyskane wartości dla poszczególnych różnią się od , przy czym tylko dla parametru (przy liczbie kart płatniczych) wartość odchylenia standardowego *a posteriori* jest niższa niż 10 (bo ). Można by zatem przypuszczać, że w ogólności (z wyjątkiem parametru ) informacja płynąca z danych zwiększyła niepewność wnioskowania o parametrach (co można by odbierać negatywnie), ALE(!) warto przyglądnąć się temu, jak mają się uzyskane wartości odchyleń standardowych *a posteriori* do wartości oczekiwanych *a posteriori*. Weźmy przykładowo następujące wyniki: i *versus* i . *Relatywnie* (w stosunku do wartości oczekiwanej) niepewność *a posteriori* parametru jest wyższa aniżeli ta dla parametru . Ponadto, wyniki dla parametru już sugerują, że gdy zbudujemy przedział HPD dla tego parametru, wówczas będzie on zawierał w sobie 0 (a to oznacza, że parametru nie można uznać za statystycznie istotny – o czym jeszcze w kolejnym punkcie, poniżej). Natomiast, dla parametru , owszem, niepewność jest wyższa (), wokół wartości oczekiwanej, która jest znacznie bardziej oddalona od 0, aniżeli ta dla , co może nam sugerować, że gdy zbudujemy przedział HPD dla , wówczas nie będzie on zawierał w sobie wartości 0, a zatem parametr ten okaże się statystycznie istotny.

Z powyższej dyskusji wynika, że odchylenia standardowe *a posteriori*:

* na początek porównujemy z tymi przyjętymi w rozkładzie *a priori*, a …
* … następnie porównujemy (wzrokowo) z wartościami oczekiwanymi *a posteriori*, by …
* … w końcu przejść do analizy przedziałów HPD, bo te najszybciej nas poinformują – uwzględniając w sobie zarówno położenie, jak i rozproszenie rozkładu *a posteriori* – o istotności poszczególnych parametrów
* Przedziały najwyższej gęstości *a posteriori* (HPD):
	+ Pod kątem interpretacji – ta jest szablonowa, weźmy na przykład przedział HPD dla :

*Po wglądzie w dane, z prawdopodobieństwem a posteriori* *równym parametr mieści się w przedziale . Ponadto, przedział ten jest najkrótszy z możliwych.*

* + Pod kątem weryfikacji (testem Lindley’a – patrz wykład) statystycznej istotności poszczególnych parametrów – tę badamy poprzez proste sprawdzenie, czy wartość 0 „wpada” czy „nie wpada” do przedziału HPD dla danego parametru. Ponieważ przedział HPD obejmuje takie wartości parametru, które są najbardziej „wspierane przez dane” (najbardziej prawdopodobne *a posteriori*), wówczas jeśli 0 jest jedną z nich, to podejście bayesowskie nie pozwala odrzucić stwierdzenia (H0), że parametr jest nieistotny (czyli że jest równy 0). W przeciwnym razie, tj. gdy wartość 0 nie mieści się w danym przedziale HPD (a zatem cały przedział jest bądź na lewo, bądź na prawo od 0), to oznacza to, że wszystkie najbardziej wspierane przez dane wartości parametru są bądź ujemne, bądź dodatnie – ale w obydwu przypadkach *różne* od 0, co prowadzi do konkluzji, że taki parametr należy uznać za statystycznie istotny (z prawdopodobieństwem *a posteriori* równym – tym, przy którym wyznaczono przedział HPD).

🡪 Uwaga: Na koniec zauważmy, że całość uwagi powyżej poświęcona została wnioskowaniu tylko o parametrach strukturalnych (czyli , ). Tymczasem w modelu obecny jest jeszcze jeden: – czyli precyzja składników losowych modelu regresji, która jest parametrem tylko zakłócającym (ang. *nuisance parameter*), o drugorzędnym znaczeniu, a tym samym nie stanowiącym „zbyt ważnego” przedmiotu naszego zainteresowania (w odróżnieniu od „bet”). Co prawda można by porównać ze sobą i oraz i (w arkuszu są już obliczone), ale już to tu sobie darujemy.