

Projekt zaliczeniowy

- 1) **Celem** projektu jest przeprowadzenie w parach prostego badania empirycznego w **arkuszu MS Excel**, poświęconego budowie i analizie bayesowskiego modelu normalnej regresji liniowej (BMNRL) z gamma-normalnym rozkładem *a priori* z zależnością pomiędzy β i τ („GNz”), z wykorzystaniem samodzielnie zebranych danych.
- 2) **Skład zespołów projektowych** – projekt jest wykonywany w parach; skład zespołów należy zadeklarować w arkuszu *Skład par projektowych / Terminarz zaliczeń*, do którego link znajduje się na stronie naszego kursu (w sekcji „Projekt zaliczeniowy”); w przypadku nieparzystej liczby osób, jedna wykonuje projekt samodzielnie, uzyskując od razu dodatkowe 10 pkt (na 100 możliwych i utożsamianych z wynikiem 100% z całości przedmiotu)
- 3) **Forma opracowania projektu** – na projekt składa się tylko 1 plik – arkusz MS Excel, w którym powinny się znajdować:
 - temat badania (przykładowo: *Bayesowska analiza zależności długości życia od liczby lekarzy i liczby szpitali w Polsce w latach 1960-1998*)
 - zwięzły opis modelowanego problemu, wraz z motywacją i doborem zmiennych
 - modelowany zbiór danych
 - opis danych:
 - czego dotyczą / co reprezentują obserwacje?
 - jakiego zakresu przestrzennego / czasowego dotyczą?
 - w jakich jednostkach są wyrażone?
 - wskazanie źródła / źródeł danych
 - jaka jest częstotliwość danych (w przypadku szeregów czasowych)?
 - ile wynosi liczba obserwacji?
 - zapis równania modelu (proszę wpisać je „ręcznie” w arkuszu – jak na ćwiczeniach – lub wykorzystać edytor równań: zakładka „Wstawianie” → „Równanie”)
 - zapis przyjętych założeń co do rozkładu *a priori*
 - wszelkie niezbędne obliczenia
 - prezentację wyników estymacji (jednak BEZ interpretacji!!! O te zapytam dopiero w trakcie zaliczenia)
- 4) W aspekcie obliczeniowo-interpretacyjnym należy się posiłkować zadaniem, które było przerabiane w czasie zajęć w MS Excel (bayesowski model regresji dla wysokości wypłat gotówkowych z bankomatów w Polsce)
- 5) **Zaliczanie projektu:**
 - Zasadniczo, zaliczenia projektów odbywają się w formie stacjonarnej (pok. 307 / paw. A), w określonych terminach, we wcześniej ustalonych godzinach, na które należy się zapisać w arkuszu *Skład par projektowych / Terminarz zaliczeń* udostępnionym Państwu na Moodle (w sekcji „Projekt zaliczeniowy”).
 - Zaliczenie w formie zdalnej wymaga otrzymania ode mnie zgody drogą mailową i będzie przysługiwało tylko w uzasadnionych przypadkach; wtedy, do zaliczenia w formie zdalnej, zostanie wykorzystany nasz wykładowy zespół na MS Teams
 - Przed przystąpieniem do zaliczenia (może być nawet w jego dniu) **proszę, aby każda z osób z danej pary projektowej przesłała wszystkie pliki na Moodle**, korzystając z modułu „Przesłanie plików” w sekcji „Projekt zaliczeniowy”
 - Na zaliczenie projektu dany zespół przybywa w pełnym składzie, przy czym obydwie osoby muszą dysponować swoim plikiem (aby w razie potrzeby, móc się przełączać przy udostępnianiu ekranu).
 - Na zaliczenie projektu (w formie stacjonarnej) każdy zespół przybywa **z własnym laptopem**.

- Zaliczenie projektu przygotowanego przez dany zespół polega na przedstawieniu i omówieniu kolejnych elementów projektu.
- W opracowywanych plikach nie należy umieszczać komentarzy czy interpretacji wyników estymacji – po prostu, należy je sobie wcześniej przemyśleć tak, aby w trakcie zaliczania projektu móc je bez problemu zwerbalizować.
- Przy zaliczaniu projektu:
 - a. każda z osób wchodzących w skład danego zespołu będzie „o coś” zapytana
 - b. obydwie osoby muszą znać szczegóły wykonania całego projektu, także w sytuacji, gdy wcześniej ustalony był między nimi podział pracy; tym samym nie dopuszcza się sytuacji, w której dana osoba nie potrafi opisać i wyjaśnić tej części projektu, która za którą nie była bezpośrednio odpowiedzialna w trakcie opracowywania projektu
 - c. istnieją dwa alternatywne warianty oceny projektu:
 - obydwu osobom przyznawana jest jednakowa liczba punktów, stanowiąca średnią arytmetyczną z liczby punktów uzyskanych przez poszczególne osoby za ich odpowiedzi,
 - każdy członek zespołu otrzymuje tę liczbę punktów, którą uzyskał tylko za swoje odpowiedzi;
 - d. po „wewnętrznej naradzie”, następującej po zakończeniu zaliczania projektu, każdy zespół sam dokonuje wyboru wariantu oceny swojego projektu.
- Maksymalna liczba punktów możliwych do uzyskania: 100 (utożsamiane z wynikiem 100% z całości przedmiotu)

6) Dane – ogólne wytyczne

- Dane mogą dotyczyć dowolnego zjawiska empirycznego z obszaru (bardzo) szeroko rozumianego życia społeczno-gospodarczego (zmienną objaśnianą mogą być, przykładowo: zyski, przychody, wielkość sprzedaży, koszty, cena jakiegoś dobra/usługi, stopa zwrotu, zmienne makroekonomiczne (np. PKB, stopa bezrobocia, inflacja, kurs walutowy), ale także np. liczba wizyt w muzeach, cena samochodu, liczba urodzin żywych, liczba zgonów z określonej przyczyny, liczba wypadków drogowych; dane mogą także dotyczyć np. rynku energii elektrycznej – czy to jej ceny, czy wielkości zużycia)
- Przykładowe źródła danych mogą Państwo znaleźć w zasobie pod tym właśnie tytułem („Przykładowe źródła danych”) w sekcji „Projekt zaliczeniowy” na stronie naszego kursu. Naturalnie, mogą Państwo wykorzystać także inne źródła.
- Dane mogą mieć charakter:
 - danych przekrojowych (inaczej: przestrzennych; pochodzących z różnych jednostek badania, np. krajów, przedsiębiorstw, banków, sklepów, pojedynczych osób, np. klientów, itp.)
 - szeregów czasowych
- Minimalna liczba obserwacji (T):
 - dla danych przestrzennych: 20
 - dla danych czasowych o częstotliwości:
 - rocznej: 20
 - kwartalnej, miesięcznej: 60
 - wyższej niż miesięczna: 200

ALE – zasadniczo – **im więcej, tym lepiej!** (a przynajmniej ciekawiej, bo tak naprawdę, nie zawsze lepiej... ;)

- Dane nie muszą dotyczyć *stricte* gospodarki Polski. Z reguły dla gospodarek zachodnich dostępne są bogatsze zbiory danych / dłuższe szeregi czasowe.
- W badaniu należy samodzielnie wybrać i wykorzystać co najmniej 3 zmienne objaśniające, posiadające potencjalny wpływ na zmienną objaśnianą (lub, ogólniej, będących w zależności z nią, jak np. stopa wzrostu PKB i stopa inflacji – tu ciężko o jednoznaczne ustalenie kierunku wpływu jednej zmiennej na drugą, gdyż pomiędzy obiema występuje pewne sprzężenie zwrotne... Jednak na poziomie modelu jednorównaniowego – a takim jest KMNRL – możemy arbitralnie „obsadzić” jedną ze zmiennych w roli zmiennej objaśnianej, a drugą – w roli zmiennej objaśniającej).
- W przypadku danych przekrojowych dotyczących grupy państw, należy określić konkretny „klucz” ich doboru. Przykładowo, analizę danego zjawiska możemy chcieć przeprowadzić dla krajów Unii Europejskiej (według jej składu na dany rok) czy też państw członkowskich strefy Schengen, albo stanów w USA. Oczywiście dane przekrojowe nie muszą dotyczyć krajów ;) (jeśli jednostkami badania będą pojedyncze przedsiębiorstwa danej branży czy klienci danego banku, wówczas też mamy do czynienia z danymi przekrojowymi – tu „kluczem” jest, odpowiednio, konkretna branża lub bank)
- W przypadku szeręgów czasowych zaleca się pozyskanie danych o możliwie najwyższej częstotliwości (np. zamiast rocznej stopy inflacji lepiej wykorzystać dane kwartalne czy miesięczne) – po pierwsze zwiększa to liczbę obserwacji, na podstawie których estymowane są modele (zatem szacunki parametrów i wnioskowanie statystyczne powinny być dokładniejsze), a po drugie może umożliwiać badanie dodatkowych zależności (np. analizę sezonowości czy wychwytywanie pojedynczych obserwacji odstających widocznych w obrębie np. miesiąca, ale „ginących” na przestrzeni dłuższego interwału czasowego, np. całego roku)
- W przypadku szeręgów czasowych zasadne może być uwzględnienie w modelu trendu liniowego (jeśli da się go „jakoś” dojrzeć na wykresie wartości zmiennej objaśnianej). Ponadto, jeśli mamy do czynienia z danymi o częstotliwości miesięcznej, kwartalnej czy dziennej (rynk energetyczne, finansowe), możliwe jest występowanie sezonowości – wówczas można (nie trzeba, ale powinno się :) uwzględnić w modelu zestaw sezonowych zmiennych zero-jedynkowych (w modelu powinien występować wyraz wolny, więc wtedy – dla uniknięcia problemu współliniowości – należy uwzględnić o jedną mniej zmiennych sezonowych niż wynosi liczba wszystkich sezonów w obrębie cyklu, pomijając dowolną spośród nich; przykładowo: dla danych miesięcznych mamy 12 sezonów, tj. miesięcy, w obrębie cyklu, tj. roku – w modelu uwzględniamy wyraz wolny oraz 11 z 12 zmiennych sezonowych, pomijając przykładowo grudzień, czy dowolny inny miesiąc)
- Być może warto rozważyć zmianę jednostek. Gdy zmienne są wyrażone w „dużych” wartościach (przykładowo PKB pojedynczego kraju może wynosić kilkadziesiąt milionów euro), wtedy lepiej przeskalać wartości tej zmiennej na wyrażone w mln euro. Oczywiście, taki zabieg znajdzie swoje odzwierciedlenie w interpretacji oceny parametru – jeśli PKB byłoby zmienną objaśniającą, to zdecydowanie bardziej zasadnym jest mówić o jednostkowym wzroście wysokości PKB jako o wzroście o 1 mln euro, zamiast o 1 euro.
- Być może warto przejść na logarytmy (niektórych) zmiennych...
- ~~NIE OBOWIĄZUJE: Być może warto przekształcić lub włączyć do modelu zmienne objaśniające w postaci nieliniowych funkcji innych zmiennych objaśniających. W tym celu warto zbudować wykres rozrzutu pomiędzy zmienną objaśnianą a wybraną zmienną objaśniającą – być może relacja pomiędzy tymi dwiema wcale nie jest liniowa, a na przykład kwadratowa czy hiperboliczna.~~
- Być może warto rozważyć przejście na wielkości względne. Niektóre zjawiska można analizować w ujęciu bezwzględnym, jak i względnym – w sensie, który najlepiej zaprezentują na przykładzie. Wyobraźmy sobie, że analizujemy wielkość sprzedaży pojazdów osobowych w różnych krajach. Te ostatnie, oczywiście, charakteryzują się (niekiedy bardzo) różnymi wielkościami populacji, przy czym naturalnie należy się spodziewać, że im wyższa liczba ludności zamieszkującej dany kraj, tym wyższa sprzedaż (co skłania do uwzględnienia w modelu wielkości populacji jako jednej ze zmiennych objaśniających). Jednakże, właśnie z uwagi na tę heterogeniczność państw

(poprzez wyraźne ich zróżnicowanie m.in. pod względem ludności) zasadne jest nierzadko przejście na *wartości względne* (tu: liczba sprzedanych pojazdów osobowych *na 100 000 mieszkańców*), które powinny się charakteryzować większą jednorodnością w porównaniu do bezwzględnych wartości. Ta większa homogeniczność obserwacji:

- po pierwsze, zapewnia porównywalność wartości danej zmiennej w różnych krajach (2000 sprzedanych samochodów w Niemczech a 2000 sprzedanych samochodów w Belgii to nie to samo, CHYBA że – no właśnie – mówimy o 2000 samochodów *na 100 000 mieszkańców*)
 - po drugie, zmniejsza ryzyko wystąpienia heteroskedastyczności składników losowych (czyli niejednorodności ich wariancji), co kwestionowałoby jedno z założeń KMNRL
- Dane należy umieścić w arkuszu MS Excel „od góry do dołu” (nie w poprzek / poziomie, jak ma to miejsce w niektórych bazach danych) – pod kątem ich dalszego bezpośredniego wykorzystania do budowy stosownych macierzy wykorzystywanych w estymacji:
 - Dane w postaci szeregów czasowych należy ułożyć tak, aby najstarsze obserwacje znajdowały się u góry arkusza, zaś „najświeższe” – na samym dole zakresu danych
 - Dane przekrojowe – co do zasady – nie muszą być uporządkowane w żaden konkretny sposób

7) Graficzna prezentacja danych

1) Wykresy prezentujące pojedyncze zmienne (objaśnianej i objaśniających):

- Dla szeregów czasowych: **wykresy liniowe**
- Dla danych przekrojowych: **wykresy słupkowe** (z wartościami poszczególnych zmiennych – objaśnianej i objaśniających – uporządkowanymi rosnąco według zmiennej objaśnianej)

2) Wykresy zależności: **wykresy rozrzutu** dla par:

(wybrana zmienna objaśniająca – na osi OX; zmienna objaśniana – na osi OY)

Uwaga: Skonstruowane wykresy należy „umieć omówić”, zwracając uwagę na widoczne wzorce, trendy, zależności, ewentualne wartości odstające (z ich hipotetycznym lub faktograficznym wyjaśnieniem ich wystąpienia). Czy kierunek zależności, które (i jeśli) dają się zauważyć na wykresach rozrzutu, jest zgodny z intuicją?

8) Część empiryczna projektu – szczegóły

- i. **Specyfikacja rozkładu *a priori*** – jak na zajęciach (choć jak najbardziej mogą też Państwo modyfikować wartości hiperparametrów)
- ii. Obliczenie odpowiednich **wielkości związanych z rozkładem *a posteriori*** – jak na zajęciach
- iii. Obliczenie **charakterystyk brzegowych rozkładów *a posteriori*** – jak na zajęciach
- iv. Skonstruowanie **wykresów gęstości brzegowych rozkładów *a posteriori*** dla poszczególnych (każdego!) parametrów strukturalnych, wraz z naniesionymi na nie gęstościami odpowiednich rozkładów *a priori*.
- v. **Dodatkowo:** Korzystając z umiejętności tworzenia wykresu gęstości *a posteriori* dla parametrów strukturalnych, mogą się Państwa pokusić także o utworzenie analogicznego **wykresu dla parametru precyzji składników losowych, τ** , który to jest stosownym rozkładem gamma. W istocie, będzie to nawet łatwiejsze, gdyż w Excelu dysponujemy gotową funkcją obliczającą wartość funkcji gęstości rozkładu gamma – należy tylko liczyć się z parametryzacją tego rozkładu, o czym akurat mowa w naszej „ściągde” z rozkładu gamma (w sekcji „Materiały pomocnicze”).
- vi. **Dodatkowo – predykcja** – jak na ostatnich zajęciach, z tym, że:
 - Jeżeli modelowane są **dane przekrojowe** („po obiektach”): proszę rozważyć $h = 2$, tj. wykonać prognozę jednocześnie dla dwóch hipotetycznych „obiektów”, samodzielnie i arbitralnie (jak na zajęciach) ustalając

wartości zmiennych objaśniających. Przykładowo, wyobraźmy sobie, że modelujemy cenę mieszkań, P_t , w zależności od powierzchni, A_t , i numeru piętra, F_t (mając dane przekrojowe, tj. „po mieszkaniach”):

$$P_t = \beta_1 + \beta_2 A_t + \beta_3 F_t + \varepsilon_t$$

Przyjęcie $h = 2$ oznacza po prostu, że prognozujemy ceny dwóch osobnych mieszkań, tak więc wektor prognoz to $y^f = (P_1^f \ P_2^f)'$. Ponieważ prognozę w ramach modeli regresji liniowej możemy dokonywać tylko przy zadanych wartościach wszystkich zmiennych objaśniających, sami musimy tu sobie zadać: metraże i numery piętra tych (hipotetycznych) mieszkań, których prognoza ceny by nas interesowała – w zupełnej dowolności, np. dla pierwszego mieszkania: $A_1^f = 50 \text{ m}^2$ i $F_1^f = 3$ piętro (wobec czego wektor „x-sów” dla tej prognozy to $x_1^f = [1 \ A_1^f \ F_1^f] = [1 \ 50 \ 3]$, zaś dla drugiego mieszkania $A_2^f = 38 \text{ m}^2$ i $F_2^f = 0$ (parter) (wobec czego $x_2^f = [1 \ A_2^f \ F_2^f] = [1 \ 38 \ 0]$).

- Jeżeli modelowane są **szeregi czasowe**: h = maksymalny horyzont prognozy możecie Państwo ustalić dowolnie (to może być prognoza na $h = 2$ okresy wprzód, ale też i dowolnie więcej); wartości zmiennych objaśniających także i w tym przypadku należy ustalić samodzielnie, tyle tylko, że – z uwagi na to, że mamy tu do czynienia z szeregami czasowymi – dobrze, żeby wartości „x-sów” w okresie prognozy „jakoś korespondowały” z ich wartościami pod koniec okresu próby (czyli tymi ostatnimi, „najnowszymi”). W szczególności można przyjąć, że w okresie prognozy (w każdym z h horyzontów) wartości „x-sów” są takie same, jak te odnotowane dla ostatniej obserwacji w próbie (tj. dla $t = T$). Można także pokusić się o jakieś bardziej subtelne podejście – biorąc pod uwagę ewentualne tendencje rozwojowe „x-sów” pod koniec okresu próby (przykładowo, jeżeli wartości danej zmiennej objaśniającej wykazywały tendencję wzrostową pod koniec okresu próby, to poza nią – czyli w okresie predykcji – można zadać tak wartości tej zmiennej, żeby „jakoś kontynuowały” tę tendencję). Ponadto, jeżeli w modelu został uwzględniony trend liniowy, czyli zmienna czasowa $t = 1, 2, \dots, T$, wówczas w okresie prognozy musi ona przyjmować kolejne (po T) wartości, tj. $t = T+1, T+2, \dots, T+h$.

- W obu przypadkach typu danych (przekrojowe vs. szeregi czasowe) proszę obliczyć charakterystyki brzegowych rozkładów predyktywnych:

- a) wartości oczekiwane,
- b) odchylenia standardowe
- c) przedziały najwyższej gęstości predyktywnej (HPredD).

Ponadto:

- Jeżeli modelowane są **dane przekrojowe**, proszę opracować wykres funkcji gęstości brzegowego rozkładu predyktywnego dla jednej z prognoz (obojętnie której)

- Jeżeli modelowane są **szeregi czasowe**, proszę:

- ALBO opracować wykres funkcji gęstości brzegowego rozkładu predyktywnego dla jednej z prognoz (obojętnie której),
- ALBO („na super extra bonus” :) opracować wykres wachlarzowy, korzystając z tego przykładowego, który udostępniłem Państwu w pliku Excel w sekcji „Ćwiczenia 3” (do samodzielnego „rozkmienienia”).