Przemysław Biecek

Analiza danych z programem R

Modele liniowe z efektami stałymi, losowymi i mieszanymi

WYDAWNICTWO NAUKOWE PWN

Analiza danych z programem R

Przemysław Biecek

Analiza danych z programem R

Modele liniowe z efektami stałymi, losowymi i mieszanymi

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 2013

Projekt okładki i stron tytułowych **Agnieszka Łydżba**

Ilustracja na okładce **serknor/Shutterstock**

Redaktor inicjujący **Berenika Grajkowska**

Redaktor **Izabela Ewa Mika**

Koordynator produkcji **Mariola Grzywacka**

Łamanie **FixPoint, Warszawa**

ISBN 978-83-01-17453-8

Wydanie II

Wydawnictwo Naukowe PWN SA
tel. 22 69 54 321; faks 22 69 54 288
infolinia 801 33 33 88
e-mail: pwn@pwn.com.pl; www.pwn.pl
Druk i oprawa: OSDW Azymut Sp z o.o

Spis treści

Przedmowa

Zanim opowiem, o czym i dla kogo jest ta książka oraz czego można z niej się nauczyć, wyjaśnię, dlaczego w ogóle ją napisałem.

Mniej więcej 10–11 lat temu wybrałem się na seminarium „Statystyka w genetyce" prowadzone przez prof. Witolda Kloneckiego w Instytucie Matematyki Politechniki Wrocławskiej. Byłem już po kilku kursach statystyki, ale z jakiegoś powodu w żaden sposób mnie ta dziedzina nie pociągała. Wyglądała jak zbiór regułek, z których trzeba wybrać odpowiednią, żeby coś policzyć, i koniec. Seminarium prof. Kloneckiego wpłynęło na całkowitą zmianę tego poglądu, statystyka stała się narzędziem do podglądania świata. Analiza danych przypomina pracę detektywa, czasami nudną, gdy trzeba godzinami przygotowywać dane i poprawić literówki lub zamienić kropki na przecinki, ale w wielu momentach fascynującą, pełną zagadek, niespodzianek i trudnych pytań. Prawdziwa analiza danych, prowadzona wspólnie z ekspertem dziedzinowym, czy to lekarzem, biologiem, czy ekonomistą, może być źródłem świetnej zabawy.

W ciągu tych 10 lat zdarzyło mi się wielokrotnie pracować z tzw. „nie-matematykami". Wielokrotnie stwierdzałem, że znacznie łatwiej byłoby mi z nimi rozmawiać, gdybym wyjaśniając, jak działa określone narzędzie statystyczne, miał pod ręką historyjkę przedstawiającą jego przykładowe wiarygodne zastosowanie. Stąd pomysł na napisanie zbioru opowiadań, które będą przedstawiały listę zagadnień. Każde opowiadanie naświetli inny problem, każde opowiadanie można przedstawić rzeczonemu „nie-matematykowi" i powiedzieć, na czym polega problem. Ta książka miała mieć podtytuł „Książka kucharska z elementami przygodowymi", tyle że nie spodobał się on recenzentom. Tytuł byłby o tyle odpowiedni, że w zamierzeniu przedstawia ona zbiór zagadnień oraz informacje jakich funkcji z programu R należy użyć, by dane zagadnienie przeanalizować, a czasami też rozwiązać. Elementy przygodowe biorą się z przykładów rzeczywistych danych zbieranych przez polskie zespoły badawcze. Zamiast uczyć się, jak rozróżniać płatki irysa, będziemy rozwiązywać problemy, które kiedyś może pozwolą na wydłużenie czasu funkcjonowania

przeszczepu u pacjenta lub umożliwią skuteczniejszą walkę z białaczką, zwiększenie produkcji mleka lub obniżenie kosztów leczenia pacjentów przy zachowaniu skuteczności.

Dlaczego akurat modele liniowe z efektami stałymi i mieszanymi? Jest to wynikiem współpracy z firmą Netezza Polska. Jakiś czas temu w czasie wakacji pomagałem przy implementacji estymatora parametrów w modelu mieszanym w równoległym środowisku. Zaskoczyło mnie wtedy, jak niewiele jest w Polsce miejsc, gdzie modele mieszane z efektami losowymi są omawiane podczas kursów statystyki, i jak inaczej wygląda to na zachodzie, gdzie znajomość modeli mieszanych przez studenta statystyki jest czymś oczywistym. Zdziwił mnie również brak polskojęzycznej literatury poświęconej aplikacjom modeli mieszanych. Stwierdziłem więc, że napiszę kilkunastostronicowe wprowadzenie do modeli mieszanych. Z biegiem czasu i wskutek różnych rozmów to wprowadzenie przerodziło się w książkę.

Do kogo adresowana jest ta książka? Napisałem coś, co sam chciałbym przeczytać kilka lat temu. Zakładam, że czytelnik ma podstawową wiedzę o statystyce i programowaniu w programie R. Zakładam też, że jest zainteresowany zastosowaniami i chce wiedzieć, kiedy jakiego narzędzia użyć, ale chce wiedzieć też, jak wykorzystywane narzędzia działają. Statystyka to nie czarna skrzynka, którą strach otworzyć, ale zbiór pomysłowych obserwacji pozwalających na kontrolowaną analizę danych.

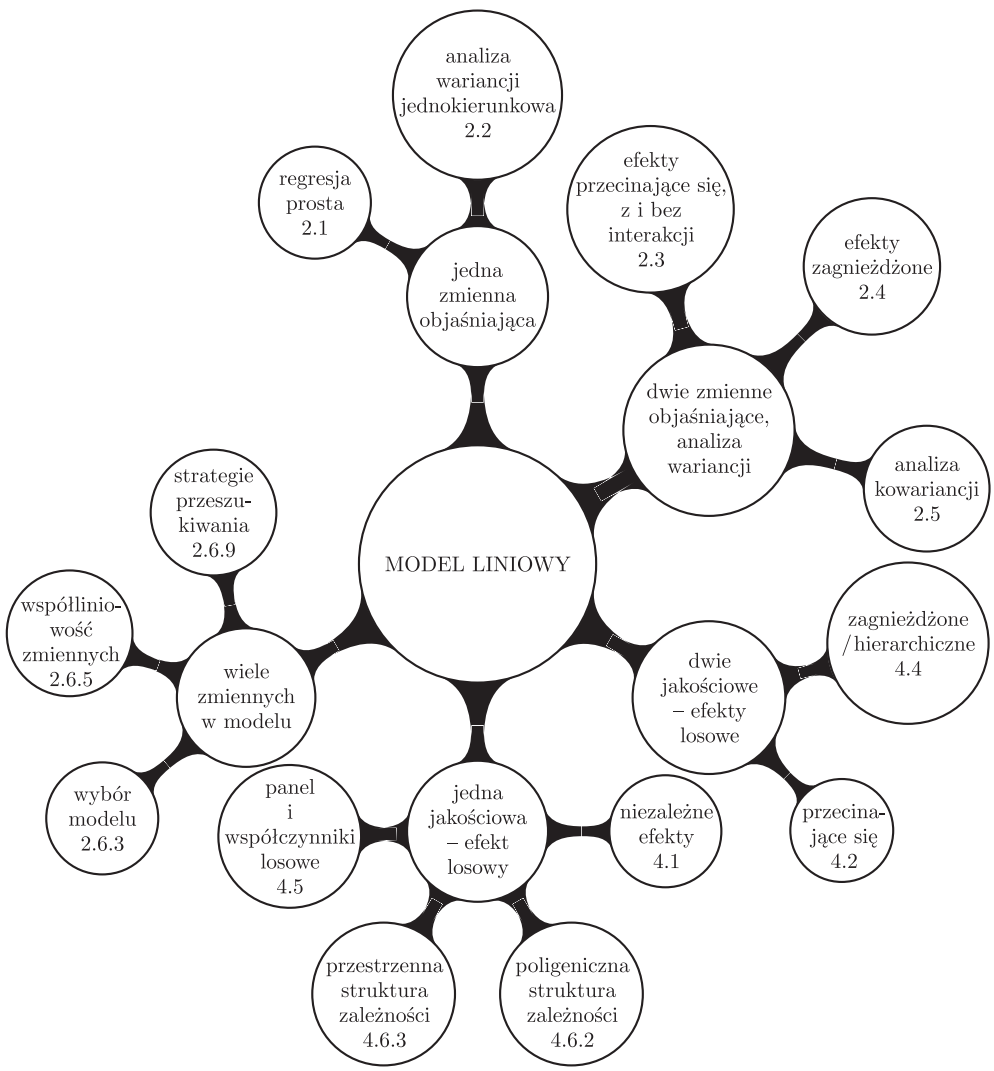
Starałem się, by ta książka była użyteczna zarówno dla bardziej zmatematyzowanych lekarzy czy biologów, jak i dla statystyków, matematyków czy informatyków, którzy lubią analizę danych. Realizacja tego planu wymagała różnych kompromisów. Nie jest to klasyczny podręcznik statystyki matematycznej. Nie wyprowadzam tutaj dowodów twierdzeń, nie przytaczam nawet twierdzeń (za wyjątkiem kilku najważniejszych). Teoria matematyczna związana z modelami liniowymi zarówno o efektach stałych, jak stałych i mieszanych jest przedstawiona w dwóch rozdziałach 1 i 3. Prezentując teorię, koncentruję się raczej na aspekcie obliczeniowym umożliwiającym stosowanie algorytmów dla dużych zbiorów danych. Osoby zainteresowane przykładami, lecz niespecjalnie zainteresowane teorią, mogą te rozdziały pominąć przy pierwszym czytaniu. Zakładam, że Czytelnik ma podstawową wiedzę o statystyce i dlatego nie tłumaczę takich pojęć jak estymator, test czy poziom istotności. Nawet jeżeli nie jest biegle zaznajomiony z tymi pojęciami, powinien nabyć intuicji z kontekstu opisywanych wyników.

One difficulty in constructing lectures or a book at this level is that most genuine data offer too many statistical problems simultaneously and often also involve biology that is unfamiliar to a first-year student. Finney, podesłane przez T.C.

Nacisk położyłem na zainteresowanie tematem i zrozumienie. Chciałbym przekonać Czytelnika, że rozróżnienie między efektem stałym a losowym ma znaczenie, że diagnostyka to rzecz potrzebna, że transformacje zmiennych to chleb powszedni, że modele konstruuje się w sposób iteracyjny. Rzadko się zdarza, że analiza danych to wykonanie testu t-Studenta w dwóch grupach albo zastosowanie modelu mieszanego do gotowego zbioru danych. Detektyw spędzi wiele godzin, czuwając w zaroślach, przemierza-

jąc kanały pod miastem, zanim pojawi się odpowiedni moment do przeprowadzenia akcji. Znaczna część praktyków uważa, że 80% czasu poświęconego na analizę danych zabiera przygotowanie danych, samo stosowanie metod statystycznych pochłania jedynie 20% czasu. Pozostała część praktyków uważa, że na przygotowanie danych trzeba poświęcić przynajmniej 90% czasu.

Na poniższym diagramie są przedstawione zagadnienia opisane w tej książce. Numery obok zagadnień to numery rozdziałów, w których są one omówione.



Wszystkie przykłady prezentowane w tej książce są wykonane z użyciem programu R. Jest to świetny, darmowy program do analiz statystycznych, zdobywający wciąż coraz większą popularność. Zakładam, że Czytelnik zna podstawy R.

W Internecie można znaleźć wiele stron z materiałami do nauki programu R, linki do wybranych materiałów umieszczone są na stronie `http://biecek.pl/R/`. W języku polskim dostępnych jest kilka książek poświęconych samemu językowi R lub też analizie danych z użyciem programu R, na przykład do poznania programu R od podstaw do zaawansowanych zastosowań może służyć książka [Biecek 2011]. Wszystkie dane, z których korzystałem w tej książce, znajdują się w opracowanym przeze mnie pakiecie `PBImisc`. Pakiet ten jest dostępny na CRAN-ie i można go zainstalować z konsoli programu R poleceniem

```
install.packages("PBImisc")
```

Na koniec chciałbym podziękować osobom, bez których ta książka by nie powstała. Przede wszystkim mojej żonie Karolinie, która po raz kolejny przymknęła oko na dziwne hobby męża, wierząc jedynie, że kiedyś się to skończy. Jestem wdzięczny Witoldowi Kloneckiemu, który jako pierwszy próbował nauczyć mnie, czym są modele mieszane. Dziękuję wielu osobom, z którymi pracowałem i które miały znaczny wpływ na moje wyobrażenie statystyki: profesorowi Stanisławowi Cebratowi, Pawłowi Mackiewiczowi (z którymi pracowałem w zakładzie Genomiki Uniwersytetu Wrocławskiego), profesor Teresie Ledwinie (pod której kierunkiem miałem przyjemność pracować w Instytucie Matematycznym Polskiej Akademii Nauk), profesorowi Jerzemu Tiurynowi (który zwerbował mnie do pracy w grupie Biologii Obliczeniowej na Uniwersytecie Warszawskim), profesorowi Janowi Mielniczukowi, profesorowi Tadeuszowi Calińskiemu, profesorowi Jackowi Koronackiemu, profesor Joannie Szydzie, profesorowi Tomaszowi Burzykowskiego, doktorom Andrzejowi Dąbrowskiemu i Andrzejowi Michalskiemu (którzy zgodzili się przeczytać wstępną bardzo surową wersję pierwszych rozdziałów). Jednocześnie muszę zaznaczyć, że nie ponoszą oni żadnej odpowiedzialności za moje ewentualne wpadki w tej książce, ponieważ wielu ich dobrych rad nie posłuchałem, mając w głowie wizję książki kucharskiej z elementami przygodowymi. Osobne serdeczne podziękowania składam Adamowi Zagdańskiemu i Arturowi Suchwałce, z którymi współpracowałem we Wrocławiu i którzy mieli znaczny wpływ na spaczenie mojego postrzegania statystyki. Chciałbym też podziękować wszystkim osobom, z którymi pracowałem nad analizą danych, w szczególności danych przedstawionych w tej książce — dużo się nauczyłem i też uświadomiłem sobie, ile jeszcze trzeba się nauczyć. Wiele też nauczyłem się, prowadząc kurs „Modele liniowe i mieszane na przykładach z biologii i medycyny" na wydziale MIMUW. Dziękuję studentom z tego kursu za wyrozumiałość i wiele uwag, którymi się podzielili, a w szczególności Oldze Kowalczuk, Aleksandrze Maj, Michałowi Lisowi i Janowi Matuszewskiemu, których krytyczne uwagi i komentarze były bardzo przydatne. Chciałbym również gorąco podziękować redaktor Izabeli Mice za włożony wysiłek w pracę nad książką.

1

Modele liniowe — wprowadzenie, podstawowe twierdzenia i wzory

1.1. Wprowadzenie

Modele liniowe to jedna z najstarszych i najpopularniejszych metod modelowania zależności między zbiorem zmiennych *objaśniających* a zmienną ilościową nazywaną zmienną *objaśnianą*. Zależność tę modeluje się zwykle po to, by móc szacować punktowo lub przedziałowo zmienną objaśnianą na podstawie zmiennych objaśniających, lub też po to, by lepiej zrozumieć zależności między obserwowanymi zmiennymi. Model liniowy można wykorzystać również do oceny, które zmienne objaśniające i w jaki sposób są zależne od zmiennej objaśnianej.

Za początek modeli liniowych uznać można prace z przełomu XVIII i XIX wieku dotyczące metody najmniejszych kwadratów, wykorzystywanej wówczas w nawigacji oraz astronomii. Nie ma jednoznacznej odpowiedzi na pytanie, kto pierwszy wymyślił tę metodę — do dziś problem pierwszeństwa rozpala dyskusję pośród historyków matematyki. Pierwsza opublikowana praca poświęcona tej metodzie pochodzi z roku 1805 i jest autorstwa francuskiego matematyka Adrien-Marie Legendre'a. Jednak Johann Carl Friedrich Gauss twierdził, że używał tej metody od ok. 1794 lub 1795 roku (Gauss miał wtedy 18 lat!) do obliczeń na potrzeby swoich prac dotyczących astronomii. W 1809 roku Gauss opisał metodę najmniejszych kwadratów w swojej książce, podając liczne przykłady zastosowań oraz udowadniając właściwości tej metody w sytuacji, gdy zakłócenie losowe ma rozkład normalny. Legendre wcale nie przyjął do wiadomości oświadczenia o pierwszeństwie Gaussa, a spory o to, kto wymyślił metodę najmniejszych kwadratów, trwają do dziś. O historii tego sporu można przeczytać np. w artykułach „C. F. Gauss and the Theory of Errors" [Sneynin 1979] lub „Gauss and the Invention of Least Squares" [Stigler 1981].

Termin *regresja* został użyty po raz pierwszy przez Sir Francisa Galtona jeszcze w XIX wieku. Francis Galton był człowiekiem renesansu, zajmującym się zarówno

antropologią, statystyką, jak i wieloma innymi naukami. W 1886 roku opublikował pracę „Regression towards mediocrity in hereditary stature" w *The Journal of the Anthropological Institute of Great Britain and Ireland* [Galton 1886], w której przedstawił wyniki badań nad dziedziczeniem wzrostu. Galton zauważył, że synowie bardzo wysokich ojców są średnio wyżsi niż synowie niższych ojców, ale ich średni wzrost jest niższy niż średni wzrost ojców. Podobnie wzrost synów niskich ojców jest bliższy średniej w populacji. To zjawisko nazwał „dążeniem do przeciętności" lub „tendencją do przeciętności" (w języku angielskim ten współczynnik nazywa się ang. *regression toward the mean*). Galton zaproponował równanie opisujące zależność między wzrostem synów i ojców lub równoważnie opisujące regresję wzrostu z pokolenia na pokolenie w kierunku wartości przeciętnej (dziś nazwalibyśmy to równanie równaniem regresji). Używając tego równania, Galton wyznaczył współczynnik dziedziczenia takiej cechy jak wzrost. Od tej pracy przyjęła się nazwa regresja na równania opisujące zależności między zmiennymi i nazwa ta upowszechniła się tak bardzo, że dziś jest stosowana nawet w sytuacjach, które nie mają żadnego związku z oryginalnym znaczeniem słowa regres.

Za słownikiem języka polskiego PWN: regres [fr. regres, łac. regressus, odejście], cofanie się w rozwoju; przeciwieństwo postępu.

Jako ciekawostkę można dodać, że zbiór danych zebrany przez Galtona jest dostępny w programie R pod nazwą `galton` w pakiecie `UsingR`. W tym zbiorze danych w jednej kolumnie jest średni wzrost rodziców liczony jako średnia wzrostu ojca i 1,08 wzrostu matki (ang. *midparent*), a w drugiej kolumnie wzrost dorosłego syna. Wartości z tego zbioru są przedstawione na rys. 1.1. Przerywana linia odpowiada prostej o równaniu $y = x$. Jak można zauważyć, w każdej grupie wzrostowej rodziców średni wzrost synów jest bliższy średniej populacyjnej.

Metody analizy modeli regresyjnych, głównie modeli o strukturze liniowej, zostały znacznie rozwinięte na początku XX wieku przez Karla Pearsona i jemu współczesnych. Dziś, wiek później, regresja jest jednym z najbardziej popularnych narzędzi modelowania statystycznego. Regresja jest popularna, ponieważ pozwala na opisanie związku między zmiennymi objaśniającymi a zmienną objaśnianą, oszacowanie średniej wartości zmiennej objaśnianej w zależności od zmiennych objaśniających, a także wybranie zmiennych istotnie zależnych od zmiennej objaśnianej. Postulowany w modelu związek między zmiennymi może mieć różnoraki charakter. Model regresji liniowej można relatywnie łatwo rozszerzyć na uogólniony model liniowy (w tej rodzinie jest bardzo popularna regresja logistyczna) lub na model nieliniowy (np. model GAM, modele krzywych sklejanych). W tej książce poświęcimy dużo uwagi modelom liniowym. Jeżeli zostaną one odpowiednio głęboko zrozumiane, to znacznie łatwiej będzie poznać bardziej elastyczne modele.

W podrozdziale 1.2 przedstawimy ogólną postać modelu liniowego, w podrozdziale 1.3 — najpopularniejsze metody estymacji współczynników w modelu liniowym, w podrozdziale 1.5 — wybrane metody testowania hipotez dotyczących współczynników w modelu liniowym oraz konstrukcji przedziałów ufności dla tych współczynników.