

# Bayesowska analiza liniowego modelu prawdopodobieństwa w prognozowaniu przynależności do grupy o wysokich dochodach na przykładzie Stanów Zjednoczonych

## 1 Cel analizy i tło teoretyczne

Celem analizy jest bayesowska estymacja modelu regresji liniowej dla danych *Adult Income*. Zbiór danych został utworzony na podstawie bazy spisowej z 1994 roku. Zgodnie z opisem danych, ekstrakcji dokonał Barry Becker, wybierając relatywnie czyste rekordy spełniające pewne warunki.

Oryginalne zadanie predykcyjne polega na określeniu, czy roczny dochód danej osoby przekracza 50 tys. dolarów. W niniejszej pracy zmienna celu zostaje zatem sprowadzona do zmiennej binarnej:

$$Y_i = \begin{cases} 1, & \text{jeżeli dochód osoby } i \text{ przekracza 50 000 USD rocznie,} \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

Tłem ekonomicznym analizy jest teoria kapitału ludzkiego oraz równanie Mincera. Klasyczny model Mincera wyjaśnia logarytm wynagrodzenia za pomocą edukacji oraz doświadczenia zawodowego, wskazując, że wyższy poziom edukacji i akumulacja doświadczenia powinny być dodatnio związane z poziomem dochodu [3, 1]. W niniejszej analizie nie obserwujemy jednak bezpośrednio ciągłej zmiennej wynagrodzenia, lecz jedynie informację, czy dochód przekracza ustalony próg. Dlatego stosowany model ma charakter uproszczony: zamiast modelować logarytm płacy, modelujemy prawdopodobieństwo przynależności do grupy wysokich dochodów.

Zastosowany zostaje *liniowy model prawdopodobieństwa* (ang. *linear probability model*, LPM), który można zapisać jako:

$$Y_i = X_i\beta + \varepsilon_i.$$

W modelu tym warunkowa wartość oczekiwana ma interpretację prawdopodobieństwa:

$$\mathbb{E}(Y_i | X_i) = P(Y_i = 1 | X_i) = X_i\beta.$$

Należy jednak podkreślić ograniczenia tego podejścia. Po pierwsze, model liniowy może generować predykcje spoza przedziału  $[0, 1]$ . Po drugie, dla zmiennej binarnej wariancja składnika losowego zależy od  $X_i$ , co oznacza heteroskedastyczność. Po trzecie, parametry należy interpretować jako zależności warunkowe, a nie jako efekty przyczynowe. W szczególności zmienne takie jak edukacja, stan cywilny czy płeć mogą być powiązane z nieobserwowalnymi cechami jednostki.

## 2 Dane i zakres próby

Wykorzystano zbiór *Adult Income*, zawierający następujące zmienne:

| Zmienna        | Opis skrótowy                             |
|----------------|-------------------------------------------|
| age            | wiek jednostki                            |
| workclass      | typ zatrudnienia / klasa pracy            |
| fnlwgt         | finalna waga obserwacji                   |
| education      | poziom edukacji jako kategoria            |
| education_num  | poziom edukacji w formie liczbowej        |
| marital_status | stan cywilny                              |
| occupation     | zawód                                     |
| relationship   | relacja w gospodarstwie domowym           |
| race           | rasa                                      |
| sex            | płeć                                      |
| capital_gain   | zyski kapitałowe                          |
| capital_loss   | straty kapitałowe                         |
| hours_per_week | liczba godzin pracy tygodniowo            |
| native_country | kraj pochodzenia                          |
| income         | zmienna dochodowa: $\leq 50K$ lub $> 50K$ |

Pierwotny zbiór treningowy zawierał 32561 obserwacji oraz 15 zmiennych. W pierwszym kroku analizę ograniczono do osób, dla których zmienna `native_country` przyjmowała wartość `United-States`. Zabieg ten zmniejsza heterogeniczność instytucjonalną i kulturową próby, ponieważ rynek pracy w różnych krajach może podlegać odmiennym mechanizmom. Po ograniczeniu danych do USA otrzymano 29170 obserwacji.

Ponieważ wymogi zadania przewidują próbę liczącą od 30 do 500 obserwacji, z ograniczonego zbioru wylosowano prostą próbę losową bez zwracania o licznosci  $N = 500$ . Struktura zmiennej dochodowej w zbiorze USA oraz w próbie przedstawia się następująco:

| Zbiór                  | $P(Y = 0)$ , dochód $\leq 50K$ | $P(Y = 1)$ , dochód $> 50K$ |
|------------------------|--------------------------------|-----------------------------|
| USA, pełny zbiór       | 0.7542                         | 0.2458                      |
| Próba losowa $N = 500$ | 0.7240                         | 0.2760                      |

Rozkłady te są zbliżone, co wstępnie wskazuje na reprezentatywność próby względem pełnego zbioru obserwacji z USA.

## 3 Przygotowanie zmiennych i specyfikacja modelu

Zmienną objaśnianą zdefiniowano jako zmienną wskaźnikową:

$$Y_i = \mathbf{1}\{\text{dochód}_i > 50,000\}.$$

Wprowadzono również następujące zmienne zero-jedynkowe:

$$\begin{aligned} D_i^{\text{male}} &= \mathbf{1}\{\text{płeć}_i = \text{mężczyzna}\}, \\ D_i^{\text{married}} &= \mathbf{1}\{\text{status małżeński}_i \in \{\text{małżeństwo}\}\}, \\ D_i^{\text{cap}} &= \mathbf{1}\{\text{capital gain}_i > 0\}. \end{aligned}$$

Zmienne ilościowe zostały wystandaryzowane:

$$\begin{aligned}\tilde{E}_i &= \frac{E_i - \bar{E}}{s_E}, \\ \tilde{A}_i &= \frac{A_i - \bar{A}}{s_A}, \\ \tilde{H}_i &= \frac{H_i - \bar{H}}{s_H},\end{aligned}$$

gdzie  $E_i$  oznacza poziom edukacji,  $A_i$  wiek, a  $H_i$  liczbę przepracowanych godzin tygodniowo. W modelu uwzględniono również składnik kwadratowy  $\tilde{A}_i^2$ , aby uchwycić nieliniowy przebieg dochodów w cyklu życia. Standaryzacja zmiennych upraszcza interpretację parametrów modelu, ponieważ sprowadza wszystkie zmienne ilościowe do wspólnej skali o średniej równej zero i odchyleniu standardowym równym jeden. Dzięki temu współczynniki regresji można interpretować jako przybliżoną zmianę prawdopodobieństwa zajścia zdarzenia przy wzroście danej zmiennej o jedno odchylenie standardowe, co umożliwia bezpośrednie porównanie siły wpływu poszczególnych czynników.

Ostatecznie rozważany model liniowy przyjmuje postać:

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 \tilde{E}_i + \beta_2 \tilde{A}_i + \beta_3 \tilde{A}_i^2 \\ &\quad + \beta_4 \tilde{H}_i + \beta_5 D_i^{\text{male}} + \beta_6 D_i^{\text{married}} + \beta_7 D_i^{\text{cap}} + \varepsilon_i,\end{aligned}$$

gdzie  $\varepsilon_i$  oznacza składnik losowy.

Dobór zmiennych wynika z teorii ekonomii pracy. Edukacja jest podstawową miarą kapitału ludzkiego. Wiek pełni funkcję przybliżenia doświadczenia oraz pozycji w cyklu życia. Kwadrat wieku pozwala uchwycić malejące przyrosty dochodu wraz z wiekiem. Liczba godzin pracy mierzy intensywność podaży pracy. Zmienna płci może odzwierciedlać różnice w strukturze wynagrodzeń i partycypacji na rynku pracy. Stan cywilny może być związany zarówno ze stabilnością gospodarstwa domowego, jak i selekcją ekonomiczną. Zmienna dotycząca dodatnich zysków kapitałowych stanowi natomiast silny sygnał statusu majątkowego.

## 4 Model bayesowski i dobór rozkładów a priori

Zakładamy klasyczny model normalny:

$$y \mid \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n),$$

ze sprzężonym rozkładem a priori typu normalny-gamma:

$$\beta \mid \sigma^2 \sim \mathcal{N}(b_0, \sigma^2 B_0), \quad \tau = \sigma^{-2} \sim \text{Gamma}(a_0, d_0).$$

Przyjęto  $a_0 = 2$  oraz  $d_0 = 1$ . Wektor średnich a priori  $b_0$  oraz odchylenia standardowe użyte do konstrukcji macierzy  $B_0$  przedstawiono w tabeli.

| Parametr            | Średnia a priori | SD a priori | Orientacyjny zakres 95% | Uzasadnienie                                        |
|---------------------|------------------|-------------|-------------------------|-----------------------------------------------------|
| $\beta_0$           | 0.25             | 0.20        | [−0.142, 0.642]         | bazowe prawdopodobieństwo wysokiego dochodu         |
| $\beta_{education}$ | 0.08             | 0.08        | [−0.077, 0.237]         | dodatni wpływ kapitału ludzkiego                    |
| $\beta_{age}$       | 0.05             | 0.08        | [−0.107, 0.207]         | dodatni wpływ doświadczenia / wieku                 |
| $\beta_{age^2}$     | −0.03            | 0.06        | [−0.148, 0.088]         | malejące przyrosty dochodu w cyklu życia            |
| $\beta_{hours}$     | 0.04             | 0.08        | [−0.117, 0.197]         | dodatni, lecz umiarkowany efekt intensywności pracy |
| $\beta_{male}$      | 0.05             | 0.10        | [−0.146, 0.246]         | możliwe różnice płacowe według płci                 |
| $\beta_{married}$   | 0.12             | 0.10        | [−0.076, 0.316]         | korelacja małżeństwa ze stabilnością ekonomiczną    |
| $\beta_{cap.gain}$  | 0.20             | 0.15        | [−0.094, 0.494]         | zyski kapitałowe jako sygnał zamożności             |

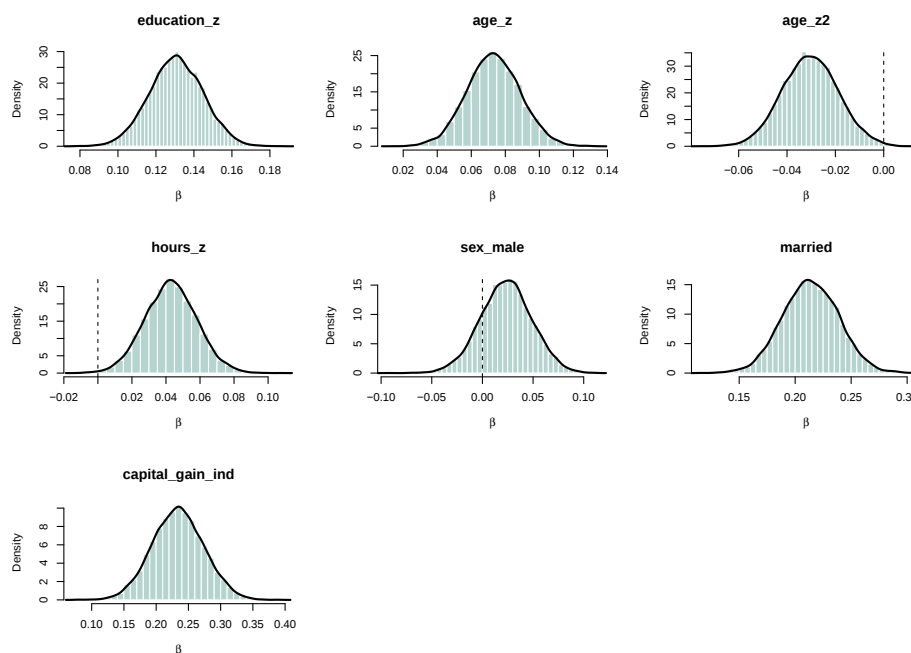
Wartości te są wynikiem elicytacji opartej na teorii kapitału ludzkiego, intuicji ekonomicznej oraz skali liniowego modelu prawdopodobieństwa. Przyjęte rozkłady a priori są umiarkowanie informacyjne: zakładają oczekiwany kierunek efektu, lecz pozostawiają znaczną swobodę danym.

## 5 Wyniki: wartości oczekiwane a posteriori

Tabela przedstawia wartości oczekiwane i odchylenia standardowe rozkładów brzegowych a posteriori parametrów.

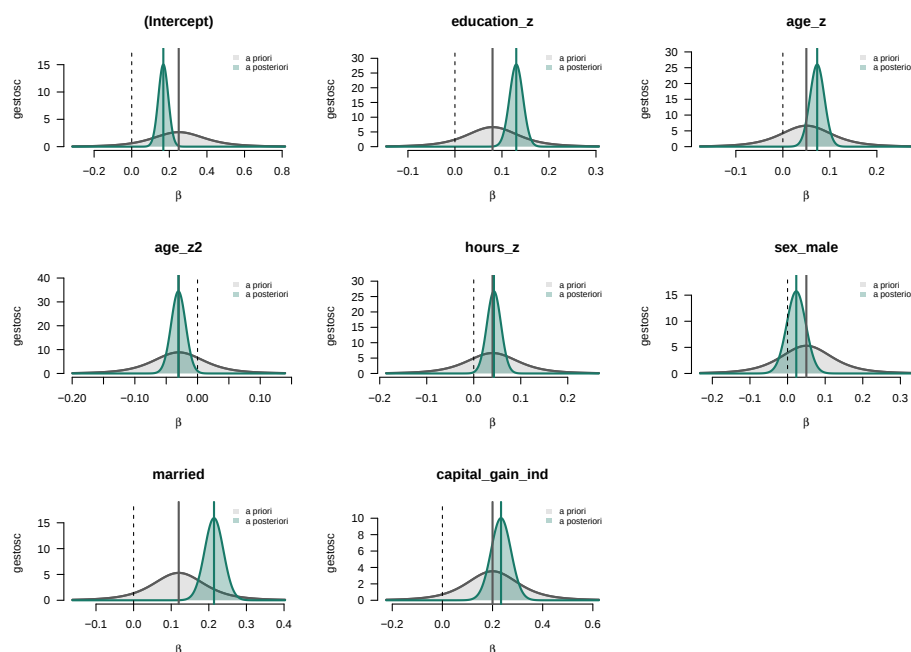
| Parametr         | Średnia a posteriori | SD a posteriori |
|------------------|----------------------|-----------------|
| Intercept        | 0.1682               | 0.0267          |
| education_z      | 0.1307               | 0.0141          |
| age_z            | 0.0731               | 0.0156          |
| age_z2           | −0.0304              | 0.0115          |
| hours_z          | 0.0428               | 0.0152          |
| sex_male         | 0.0233               | 0.0254          |
| married          | 0.2138               | 0.0251          |
| capital_gain_ind | 0.2338               | 0.0400          |

Wyniki są zgodne z intuicją ekonomiczną. Wzrost poziomu edukacji o jedno odchylenie standardowe zwiększa przewidywane prawdopodobieństwo dochodu powyżej 50 tys. USD o około 13.1 punktu procentowego. Wiek ma dodatni wpływ, natomiast ujemny współczynnik przy kwadracie wieku wskazuje na malejące przyrosty prawdopodobieństwa wraz z wiekiem. Liczba godzin pracy tygodniowo ma dodatni, umiarkowany efekt. Najsilniejsze dodatnie efekty występują dla zmiennych `married` oraz `capital_gain_ind`. Oznacza to, że osoby pozostające w małżeństwie oraz osoby z dodatnimi zyskami kapitałowymi charakteryzują się wyraźnie większym prawdopodobieństwem przynależności do grupy wysokich dochodów. Efekt zmiennej `sex_male` jest dodatni, lecz relatywnie niewielki.



Rysunek 1: Rozkłady brzegowe a posteriori parametrów modelu.

Rozkłady brzegowe a posteriori pozwalają ocenić nie tylko punktowe wartości parametrów, lecz również niepewność związaną z ich oszacowaniem. Rozkłady przesunięte wyraźnie poza zero wskazują na silną informację danych o znaku i skali danego efektu.



Rysunek 2: Porównanie rozkładów a priori i a posteriori parametrów.

W większości przypadków rozkłady a posteriori są wyraźnie bardziej skoncentrowane niż rozkłady a priori, co świadczy o dużej sile informacyjnej danych. Jednocześnie średnie a posteriori pozostają w wielu przypadkach zgodne z kierunkiem przyjętych przekonań a priori.

## 6 HPDI i interpretacja zmiennych

Tabela zawiera 95% przedziały największej gęstości a posteriori (HPDI) oraz prawdopodobieństwa znaku parametrów.

| Parametr         | Średnia | HPDI dolny | HPDI górny | $P(\beta > 0   y)$ | $P(\beta < 0   y)$ |
|------------------|---------|------------|------------|--------------------|--------------------|
| Intercept        | 0.1682  | 0.1139     | 0.2190     | 1.0000             | 0.0000             |
| education_z      | 0.1307  | 0.1030     | 0.1578     | 1.0000             | 0.0000             |
| age_z            | 0.0731  | 0.0431     | 0.1036     | 1.0000             | 0.0000             |
| age_z2           | -0.0304 | -0.0529    | -0.0077    | 0.0035             | 0.9965             |
| hours_z          | 0.0428  | 0.0133     | 0.0732     | 0.9973             | 0.0027             |
| sex_male         | 0.0233  | -0.0258    | 0.0740     | 0.8214             | 0.1786             |
| married          | 0.2138  | 0.1650     | 0.2624     | 1.0000             | 0.0000             |
| capital_gain_ind | 0.2338  | 0.1555     | 0.3117     | 1.0000             | 0.0000             |

Przedziały HPDI dla zmiennych `education_z`, `age_z`, `hours_z`, `married` oraz `capital_gain_ind` leżą w całości powyżej zera, co wskazuje na dodatni wpływ tych zmiennych na prawdopodobieństwo przekroczenia progu dochodu. Przedział HPDI dla `age_z2` znajduje się poniżej zera, co potwierdza nieliniowy profil dochodów w cyklu życia. Zmienna `sex_male` ma dodatnią średnią a posteriori, lecz jej przedział HPDI obejmuje zero. W tym przypadku dane nie dostarczają jednoznacznego potwierdzenia dodatniego wpływu płci męskiej po uwzględnieniu pozostałych zmiennych modelu.

## 7 Czynniki Bayesa

Znaczenie poszczególnych zmiennych oceniono również przy pomocy czynników Bayesa. Dla hipotezy punktowej

$$H_0 : \beta_j = 0$$

zastosowano tożsamość Savage’a-Dickey’ego:

$$BF_{01} = \frac{p(\beta_j = 0 | y)}{p(\beta_j \neq 0 | y)}, \quad BF_{10} = \frac{1}{BF_{01}}.$$

Oznacza to, że porównujemy gęstość rozkładu a posteriori i a priori w punkcie zerowym. Jeżeli po uwzględnieniu danych gęstość przy zerze spada, to dane przesuwają masę rozkładu od hipotezy  $\beta_j = 0$ , co zwiększa  $BF_{10}$ .

Do interpretacji przyjęto skalę Kass’a i Raftery’ego [2]:

| Wartość BF | Interpretacja         |
|------------|-----------------------|
| 1–3        | ledwie warte wzmianki |
| 3–20       | pozytywne wsparcie    |
| 20–150     | silne wsparcie        |
| > 150      | bardzo silne wsparcie |

| Parametr         | $BF_{01}$               | $BF_{10}$              |
|------------------|-------------------------|------------------------|
| Intercept        | $9.2260 \cdot 10^{-8}$  | $1.0839 \cdot 10^7$    |
| education_z      | $1.8796 \cdot 10^{-16}$ | $5.3203 \cdot 10^{15}$ |
| age_z            | $8.9581 \cdot 10^{-5}$  | $1.1163 \cdot 10^4$    |
| age_z2           | $1.6368 \cdot 10^{-1}$  | 6.1093                 |
| hours_z          | $8.8235 \cdot 10^{-2}$  | 11.3334                |
| sex_male         | 2.6171                  | 0.3821                 |
| married          | $1.8544 \cdot 10^{-14}$ | $5.3925 \cdot 10^{13}$ |
| capital_gain_ind | $7.5416 \cdot 10^{-7}$  | $1.3260 \cdot 10^6$    |

Czynniki Bayesa wskazują na bardzo silne wsparcie dla zmiennych `education_z`, `age_z`, `married` oraz `capital_gain_ind`. W przypadku `age_z2` oraz `hours_z` wsparcie jest pozytywne, ale słabsze. Dla zmiennej `sex_male` otrzymano  $BF_{10} < 1$ , co oznacza, że dane relatywnie bardziej wspierają hipotezę braku efektu niż hipotezę efektu różnego od zera. Jest to spójne z przedziałem HPDI obejmującym zero.

## 8 Prognoza dla wybranej obserwacji

Ostatnim elementem analizy jest predykcja dla wybranej obserwacji z próby. Dla tej obserwacji prawdziwa wartość zmiennej dochodowej wynosiła:

$$y_{true} = 1.$$

Posteriorowa prognoza wartości oczekiwanej, interpretowana w LPM jako przewidywane prawdopodobieństwo przekroczenia progu 50 tys. USD, wyniosła:

$$\hat{P}(Y = 1 \mid X, y) = 0.5243.$$

Przedział HPDI dla wartości oczekiwanej wyniósł:

$$[0.4450, 0.6011].$$

Po ograniczeniu do przedziału  $[0, 1]$  wartości pozostają bez zmian, co oznacza, że prognoza wartości oczekiwanej jest interpretowalna jako prawdopodobieństwo.

Dodatkowo obliczono pełny rozkład predykcyjny, uwzględniający składnik losowy:

$$\hat{y}_{pred} = 0.5218, \quad HPDI_{pred} = [-0.1608, 1.2405].$$

Ten szerszy przedział wychodzi poza zakres  $[0, 1]$ , co jest naturalną konsekwencją zastosowania liniowego modelu prawdopodobieństwa z normalnym składnikiem losowym. Dlatego w interpretacji predykcyjnej większy sens ma przedział HPDI dla wartości oczekiwanej  $E(Y \mid X, y)$  niż pełny przedział dla obserwacji z losowym błędem.

## 9 Podsumowanie

Przeprowadzona analiza wskazuje, że prawdopodobieństwo osiągnięcia dochodu powyżej 50 tys. USD jest dodatnio związane przede wszystkim z poziomem edukacji, wiekiem, pozostawaniem w małżeństwie oraz występowaniem dodatnich zysków kapitałowych. Ujemny parametr przy kwadracie wieku potwierdza nieliniowy przebieg relacji między wiekiem a dochodem. Wyniki uzyskane przy pomocy HPDI oraz czynników Bayesa są spójne: najsilniejsze wsparcie danych

uzyskano dla edukacji, stanu cywilnego i zysków kapitałowych, natomiast efekt płci po kontroli pozostałych zmiennych okazał się niejednoznaczny.

Zastosowany model ma charakter opisowy i predykcyjny, nie przyczynowy. Jego zaletą jest możliwość wykorzystania analitycznego rozwiązania bayesowskiego z rozkładem normalnym-gamma oraz czytelna interpretacja parametrów. Ograniczeniem pozostaje natomiast binarny charakter zmiennej zależnej, który w pełniejszym podejściu mógłby uzasadniać zastosowanie modelu logitowego lub probitowego. Wymagałoby to jednak metod numerycznych lub symulacyjnych zamiast prostego rozwiązania analitycznego.

## Literatura

- [1] Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.
- [2] Kass, R. E., Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- [3] Mincer, J. (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research.