Байєсівське усереднення для оцінки параметрів лінійної регресії

Виконав

студент 4-го курсу

групи КА-33 ІПСА «КПІ ім. І. Сікорського»

Лубинець Юрій

Керівник

к. ф.-м. н., доцент Каніовська І. Ю.

Вступ

Об’єкт дослідження: соціально-економічні дані.

Предмет дослідження: застосування байєсівського модельного усереднення для дослідження соціально-економічних даних.

Мета дослідження: проаналізувати предмет дослідження та дослідити ефективність застосування байєсівського модельного усереднення.

Методи дослідження: методи усередненої байєсівської моделі, метод біноміальної апріорної моделі, методи Монте-Карло для маркових ланцюгів.

Актуальність

Ідеї байєсівської статистики знайшли широке застосування в багатьох прикладних областях: в інженерній справі, економіці, страхуванні, медицині та інших.

Невідомий параметр ймовірнісної моделі розглядається як випадкова величина (в класичному підході цей параметр вважається фіксованою точкою).

Така інтерпретація параметра моделі дозволяє:

- апріорно наявну інформацію про невідомий параметр висловити у вигляді його апріорного розподілу;
- об'єднати апріорний розподіл за допомогою теореми Байєса з інформацією, укладеною в експериментально отриманій вибірці.

Математична постановка класичної задачі

Нехай дана лінійна модель:

$$y = \alpha_\gamma + X_\gamma \beta_\gamma + \varepsilon$$

y – залежна змінна; α_γ – константа; β_γ – коефіцієнти регресії;

ε – випадкова похибка, $\varepsilon \sim N(0; \sigma^2 I)$.

Проблематика:

- $X_j \in \{X\}$ (потенційні пояснюючі змінні) **потрібно включити в модель**?
- Наскільки вони **важливі** (за ваговими коефіцієнтами)?
- Чи потрібно включати всі змінні в лінійну модель **неефективно** та **майже неможливо** за обмеженого числа спостережень?

Математична постановка задачі байєсівського усереднення.

Модель оцінюється за її **середнім показником апостеріорної модельної ймовірності**, яка випливає із теореми Байєса:

$$p(M_j|y,\mathrm{X}) = \frac{p(y|M_j,\mathrm{X})p(M_j)}{p(y|\mathrm{X})} = \frac{p(y|M_j,\mathrm{X})p(M_j)}{\sum_{k=1}^{2^K} p(y|M_k,\mathrm{X})p(M_k)}$$

$p(y|\mathrm{X})$ – узагальнена ймовірність – константа для всіх моделей.

Розподіл зваженої апостеріорної моделі для будь-якої статистики θ (наприклад, коефіцієнта β) дорівнює:

$$p(\theta|y,X) = \sum_{\gamma=1}^{2^K} p(\theta|M_\gamma,y,\mathrm{X})p(M_\gamma|X,y)$$

Апріорна g Зельнера

Статистичним стандартом для використання «Байєсівської регресії» лінійної моделі із специфічною апріорною структурою вважають **«апріорна g Зельнера»**.

Гіперпараметр g показує наскільки точно дослідник впевнений, що **коефіцієнти близькі до нуля**.

Найбільш відомим є підхід «за замовченням» - це блок апріорної інформації, який встановлює $g = N$ загальним для усіх моделей і, таким чином, атрибути про ту ж саму інформацію містяться як спостереження.

Опис вхідних даних (1/2)

Вихідний набір даних, взятий на UCI Machine learning repository.

Модель розглянута на прикладі розглядає **залежність рівня злочинності** в районах США від **соціально-економічних показників** цих районів.

Цільова змінна, **рівень злочинності**, була взята як сума рівня жорстоких та нежорстоких злочинів.

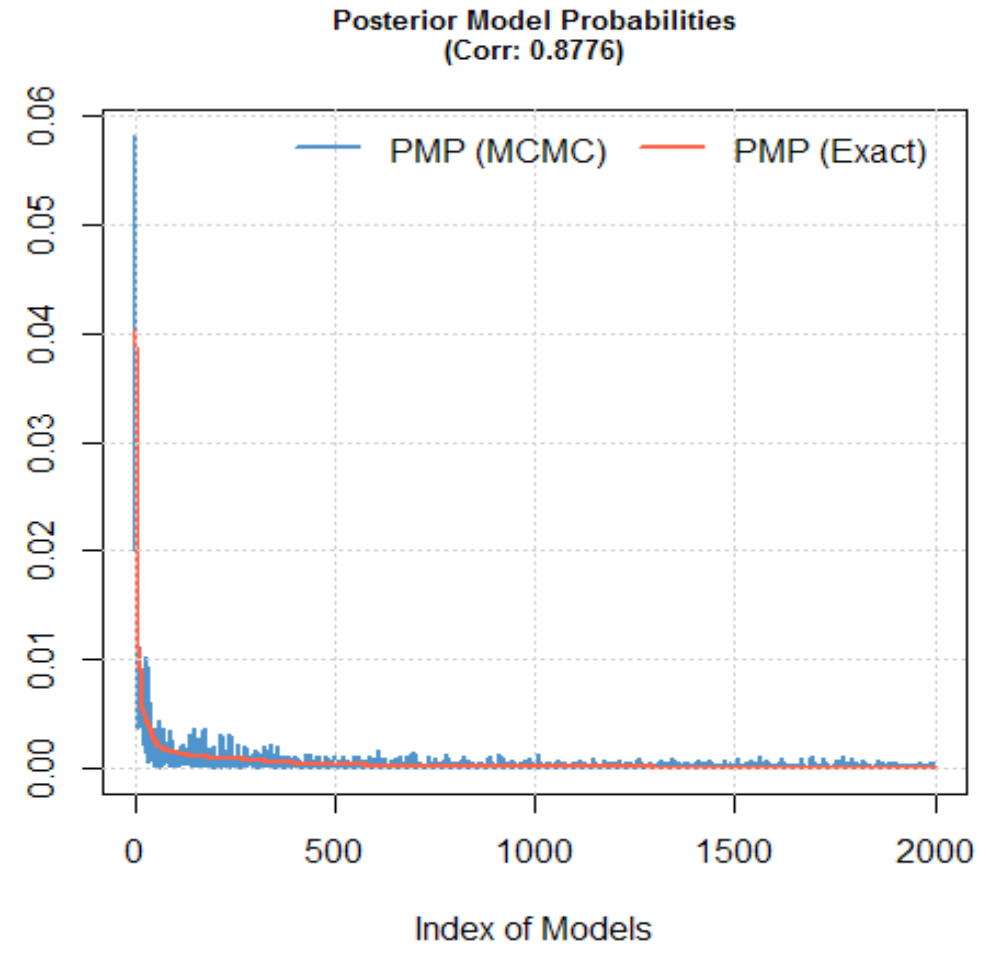
Всього в наборі даних **2215 спостережень**, зокрема 125 змінних, які можна передбачати, 4 змінні, які неможливо передбачати, 18 можливих критеріїв злочину.

Опис вхідних даних (2/2)

Апріорний коефіцієнт $g = max(N,K^2)$. Ідея полягає у тому, що апостеріорні ймовірності моделей асимптотично ведуть себе як **байєсівський інформаційний критерій** (із $g=N$), або **критерій ризика інфляції** (із $g=K^2$).

Всі числові дані були **нормалізовані**, прогнозувалася ViolentCrimesPerPop (всі інші атрибути злочину були усунені).

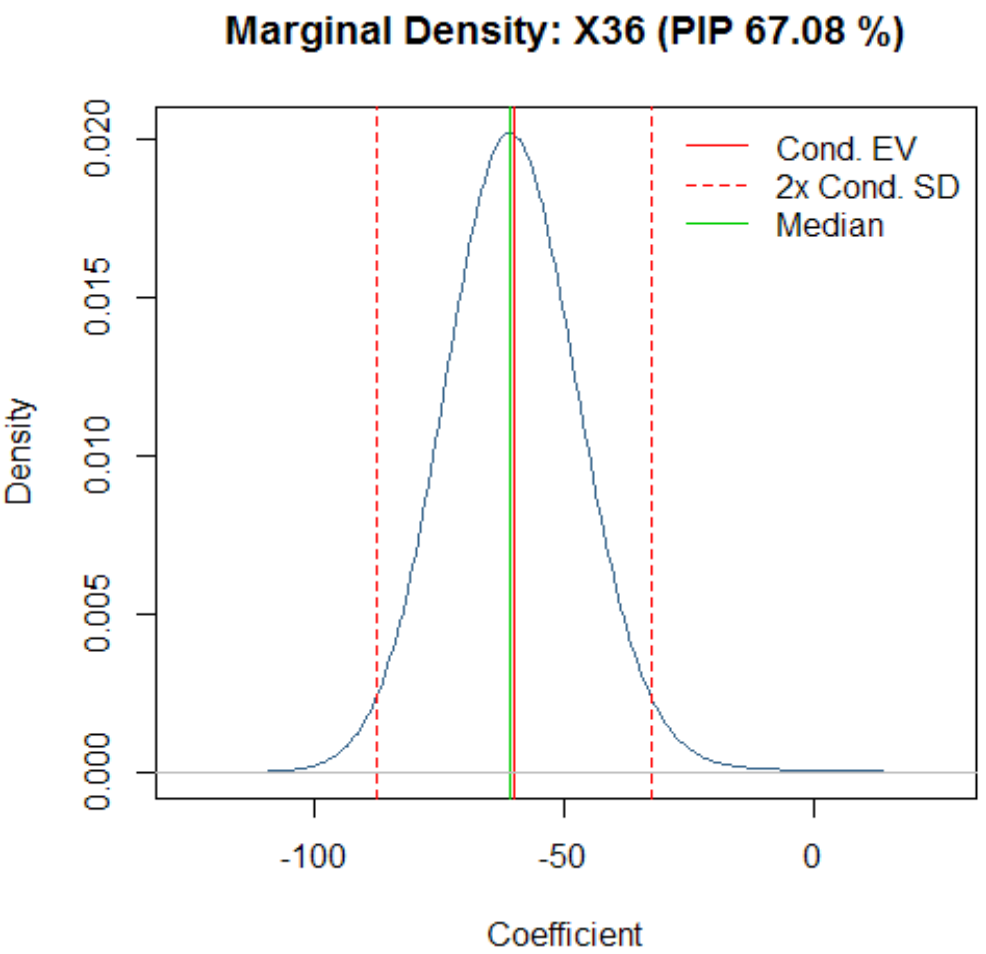
Практична реалізація



Значення **аналітичних моделей апостеріорної ймовірності** (червоний колір, PMP(Exact)) значно відрізняється від **згенерованих на марківських ланцюгах** (синій колір, PMP(MCMC)),

тому що апостеріорні модельні ймовірності, згенеровані на марківських ланцюгах, випливають із ряда ітерацій, в той час як «точні» моделі апостеріорної ймовірності підраховуються.

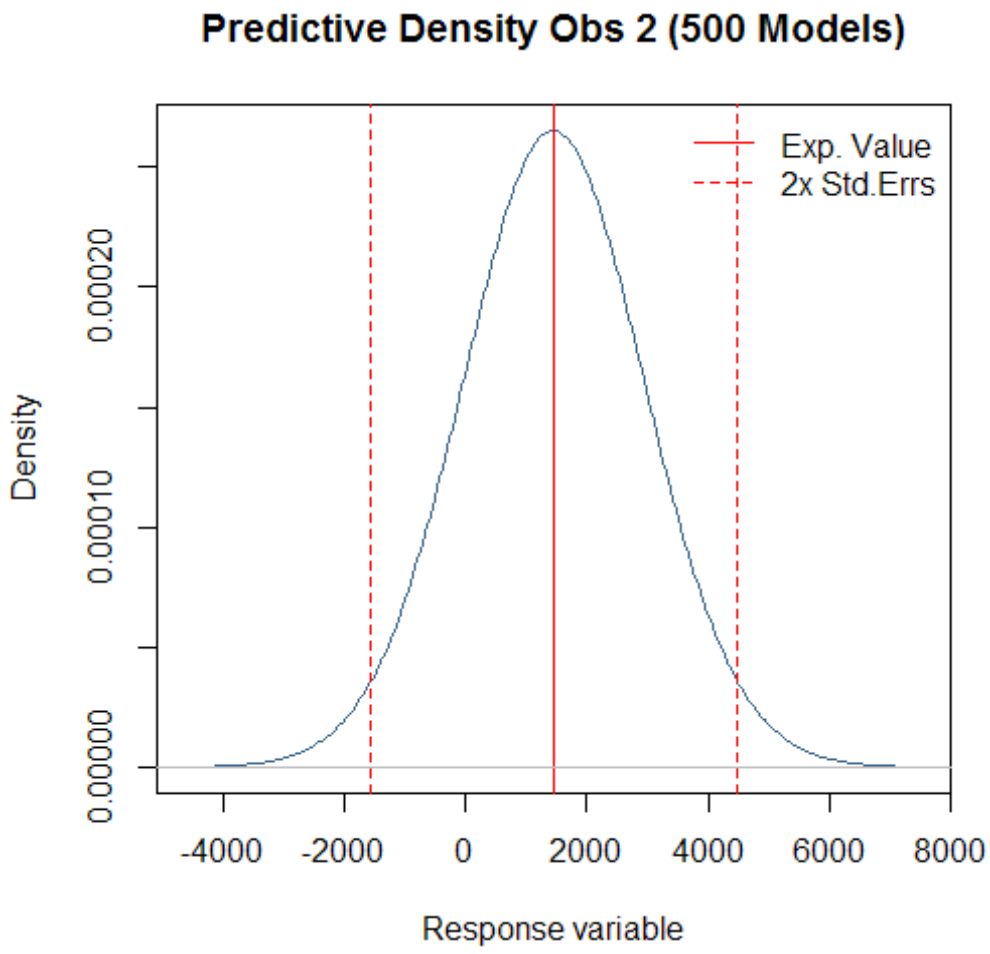
Практична реалізація



При обробці **математичного сподівання**, потрібно дослідити також всі **апостеріорні розподіли коефіцієнтів**.

Розглянемо апостеріорні щільності показника X36, (процент кількості людей, чиї доходи нижчі прожиткового мінімуму).

Практична реалізація



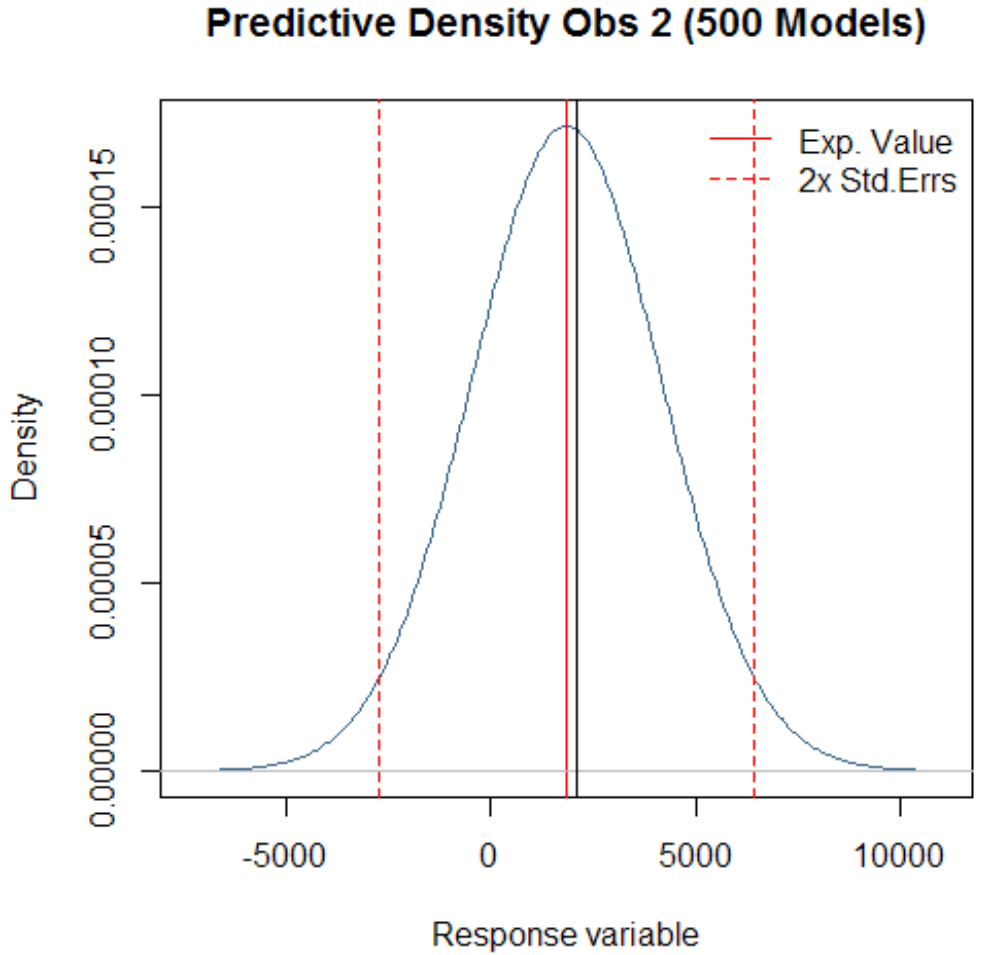
Використаємо інформацію для прогнозування злочинності різних районів США:

- BH = BerkeleyHeightstownship;
- MT = Marpletownship.

Ми бачимо, що в залежності від пояснюючих даних **очікувана злочинність** в Marpletownship дорівнює **1849.44**. А **фактичне значення** - **2083.51**.

Для порівняння моделі із фактичним результатом розглянемо прогноз помилки.

Практична реалізація



Для візуалізації прогнозу Marpletownship, порівняємо графік прогнозованої щільності з фактичним результатом. Бачимо, що **прогноз не досконалий**.

Результати показуюсь, що наша прогнозована модель не може бути виконана якісно.

Висновки

- Розглянутий метод **будує регресійні моделі** на основі байєсівської статистики.
- Цей метод **швидший** та **менш затратний** ніж стандартні методи типу марківських ланцюгів Монте-Карло.
- За даними експерименту **встановлено зв'язок** між **станом злочинності** та **характеристиками районів проживання**, що може бути корисним при прогнозуванні соціальної сфери країни, профілактики та запобігання злочинності.

Подальші дослідження

Розробка методів байєсівського модельного усереднення сприяє подальшому розширенню області практичного застосування байєсівських методів.

Можливе застосування концепції байєсівського модельного усереднення, які тут не розглядаються, такі як:

- **специфікація апріорних модельних ймовірностей**;
- **усереднення** більш ніж одного **класу моделей**;
- **інтерпретація результатів**;
- **перевірка** моделей **на адекватність**.

Дякую за увагу!