

Байесовская линейная регрессия

Максим Кочуров

МГУ им. М.В. Ломоносова

Лекция 6



- ① Байесовская интуиция
 - Эконометрика
 - Общий случай
 - Обобщенные линейные модели
 - Ограничения
- ② Байесовская линейная регрессия
 - Классические предположения
- ③ R2D2M2CP
 - Обсуждение
 - Априорное распределение R^2
 - Важность переменных
 - R2D2M2
 - Вероятность корреляции
- ④ Продвинутое обобщенные модели
- ⑤ Заключение

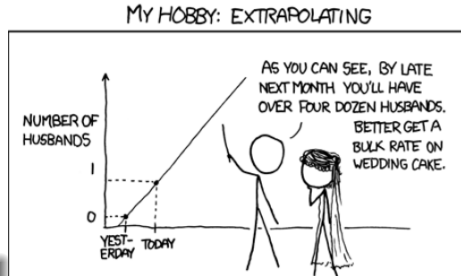


Почему линейная регрессия – это база

- Применяется при принятии решений
 - Корреляция
 - Направление влияния
 - Оценка эффектов
- Основа более сложных моделей
 - Marketing Mix Models
 - AB тесты

Lego

Линейная регрессия – основа многих статистических методов





Вводим обозначения

В эконометрике общеприняты следующие обозначения

$$y \sim x_1 + x_2 + \dots + x_k$$

Как читаем

Переменная y линейно зависит от x_1, x_2, \dots, x_k



Вводим обозначения

В эконометрике общеприняты следующие обозначения

$$y \sim x_1 + x_2 + \dots + x_k$$

Как читаем

Переменная y линейно зависит от x_1, x_2, \dots, x_k

Также предполагаем (если не оговорено иное) наличие константы

$$y \sim 1 + x_1 + x_2 + \dots + x_k$$



Вводим обозначения

В эконометрике общеприняты следующие обозначения

$$y \sim x_1 + x_2 + \dots + x_k$$

Как читаем

Переменная y линейно зависит от x_1, x_2, \dots, x_k

Также предполагаем (если не оговорено иное) наличие константы

$$y \sim 1 + x_1 + x_2 + \dots + x_k$$

Основная задача – оценка вектора коэффициентов β

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$



Нелинейные случаи

% изменение в x_1 приводит к % изменению y

$$\log y \sim \log x_1 + \dots$$



Нелинейные случаи

% изменение в x_1 приводит к % изменению y

$$\log y \sim \log x_1 + \dots$$

% изменение в x_1 изменяет y на $y.e.$

$$y \sim \log x_1 + \dots$$



Нелинейные случаи

% изменение в x_1 приводит к % изменению y

$$\log y \sim \log x_1 + \dots$$

% изменение в x_1 изменяет y на у.е.

$$y \sim \log x_1 + \dots$$

изменение x_1 на у.е. приводит к % изменению в y

$$\log y \sim x_1 + \dots$$



Нелинейные случаи

% изменение в x_1 приводит к % изменению y

$$\log y \sim \log x_1 + \dots$$

% изменение в x_1 изменяет y на у.е.

$$y \sim \log x_1 + \dots$$

изменение x_1 на у.е. приводит к % изменению в y

$$\log y \sim x_1 + \dots$$

Вывод

Интерпретация требует аккуратности в случае нелинейных связей



Обобщенные линейные модели: основы

Связать наблюдения можно с помощью произвольной функцией правдоподобия. Классический пример

$$c_i \sim \text{Binom}(p_i, n_i)$$

$$\text{link}^{-1}(p_i) \sim x_{1i} + x_{2i} + \dots + x_{ki}$$



Гетероскедастичность

Добавим немного гибкости в модель

$$y_i \sim \mathcal{N}(m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i$$



Гетероскедастичность

Добавим немного гибкости в модель

$$y_i \sim \mathcal{N}(m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i$$

Заметка

Заметим, что s_i зависит от z_i . Такие модели требуют оптимизационных методов для оценки.



Другие функции правдоподобия

Можно добиться большей гибкости, изменив функцию правдоподобия и ослабив некоторые ограничения

$$y_i \sim \mathcal{T}(\nu_i, m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i + \dots$$

$$\log \nu_i \sim w_i + \dots$$



Другие функции правдоподобия

Можно добиться большей гибкости, изменив функцию правдоподобия и ослабив некоторые ограничения

$$y_i \sim \mathcal{T}(\nu_i, m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i + \dots$$

$$\log \nu_i \sim w_i + \dots$$

Заметка

Выше – модель на основе распределения Стюдента с переменным числом степеней свободы. Её оценки очень зашумлены, если не использовать регуляризацию

Оценка модели



Из эконометрики знаем, что с помощью ММП получаем оценку

$$\hat{\beta} = (X^T X)^{-1} X^T y$$



Оценка модели

Из эконометрики знаем, что с помощью ММП получаем оценку

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- ❶ А если какие-то коэффициенты точно положительны?
- ❷ А если какие-то коэффициенты в β маленькие по величине?
- ❸ А если какие-то переменные не имеют статзначимости?

Ограничение

В частотной статистике нельзя никак применять имеющиеся знания



Априорные распределения

Байесовский подход предполагает наличие априорных распределений, но каких?

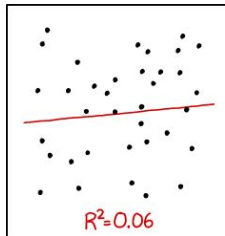
$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$

$$\beta_j \sim ???$$

$$c \sim ???$$

$$\sigma \sim ???$$

Ранее ввели два параметра: β , σ



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.



Выбираем распределения

В качестве априорного распределения возьмем нормальное распределение

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$

$$\beta_j \sim \mathcal{N}(0, 1)$$

$$c \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{N}_+(1) \quad // \text{ Half Normal}$$

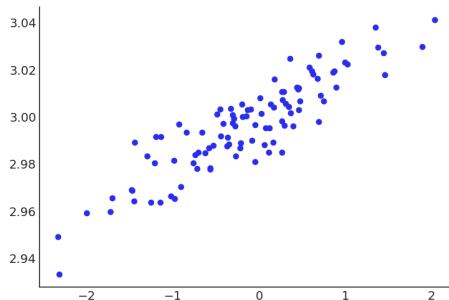


Рис.: Пример данных



Выбираем распределения

В качестве априорного распределения возьмем нормальное распределение

$$y_i \sim \mathcal{N}(c + \beta^T x_i, \sigma)$$

$$\beta_j \sim \mathcal{N}(0, 1)$$

$$c \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{N}_+(1) \quad // \text{ Half Normal}$$

Внимание

Базовые параметры для априорных β и σ могут приводить к непредсказуемым результатам

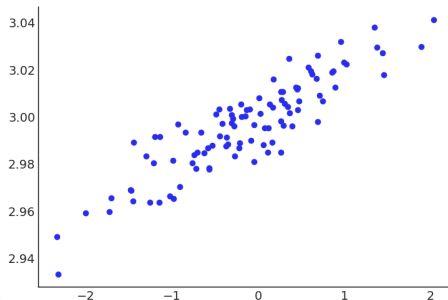


Рис.: Пример данных



Выбираем распределения

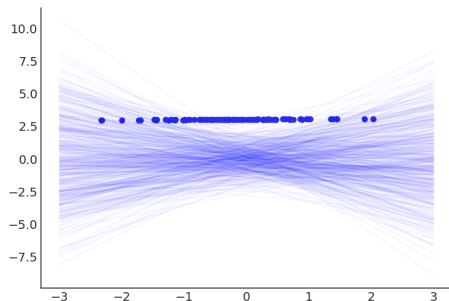
Для подбора априорного – воспользуемся предиктивным распределением

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$

$$\beta_j \sim \mathcal{N}(0, 1)$$

$$c \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{N}_+(1) \quad // \text{ Half Normal}$$



Внимание

Иногда априорное предиктивное распределение "съезжает", поэтому надо проверять визуализацию

Рис.: Априорное предиктивное распределение



Выбираем распределения

Для подбора априорного – воспользуемся предиктивным распределением

$$y_i \sim \mathcal{N}(c + \beta^T x_i, \sigma)$$

$$\beta_j \sim \mathcal{N}(0, 100)$$

$$c \sim \mathcal{N}(0, 100)$$

$$\sigma \sim \mathcal{N}_+(1) \quad // \text{ Half Normal}$$

Внимание

Иногда априорное предиктивное распределение "съезжает", поэтому надо проверять визуализацию

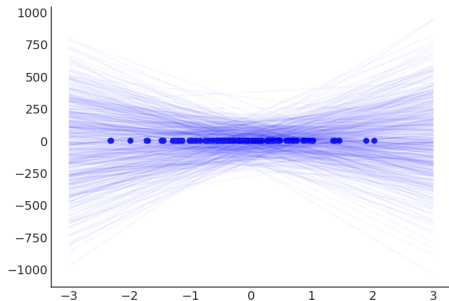


Рис.: Априорное предиктивное распределение



Больше параметров – больше проблем

Предполагая априорно независимость объясняющих переменных, можно вычислить теоретическую дисперсию

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$

$$V[y_i] = \sum V[x_{ij}] * V[\beta_j]$$

Разница в значениях в случае $x_i \in R^3$ и в случае $x_i \in R^{100}$ велика

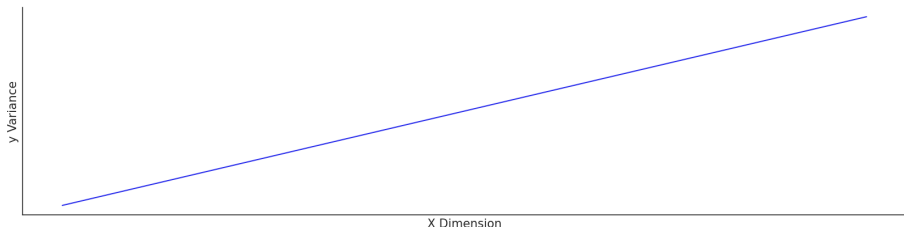


Рис.: Больше объясняющих переменных – больше дисперсия



Как это легко поправить

Простой способ убрать зависимость от числа регрессоров следующий

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$

$$\beta_j \sim \mathcal{N}(0, \frac{\sigma_\beta^2}{D})$$

...

Совет

Стандартизируйте данные: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$



Практический подход

Для начала – стандартизируем данные: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$

$$\bar{y}_i \sim \mathcal{N}(c + \bar{\beta}^\top \bar{x}_i, \sigma)$$

$$\bar{\beta} \sim \mathcal{N}(0, 1/D)$$

$$c \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{N}_+(1)$$

- 1 Дисперсия входных/выходных данных равна 1
- 2 Среднее входных/выходных данных равно 0
- 3 Почти всегда работает
- 4 Сложно подобрать априорное для σ



Практический подход

Для начала – стандартизируем данные: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$

$$\bar{y}_i \sim \mathcal{N}(c + \bar{\beta}^\top \bar{x}_i, \sigma)$$

$$\bar{\beta} \sim \mathcal{N}(0, 1/D)$$

$$c \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \mathcal{N}_+(1)$$

- 1 Дисперсия входных/выходных данных равна 1
- 2 Среднее входных/выходных данных равно 0
- 3 Почти всегда работает
- 4 Сложно подобрать априорное для σ

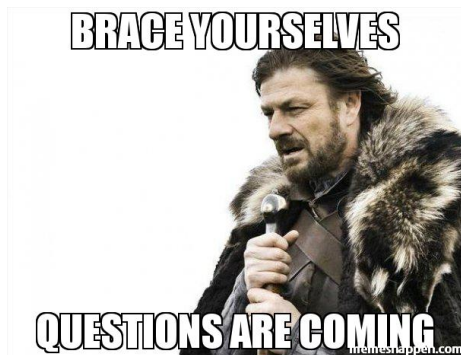
Восстановление исходных параметров

$$\beta_j = \frac{\bar{\beta}}{\text{std}(x_j)}$$

Мы знаем что мы знаем



Подбирать априорные – тяжело,
как сделать проще? Задавать себе
вопросы!

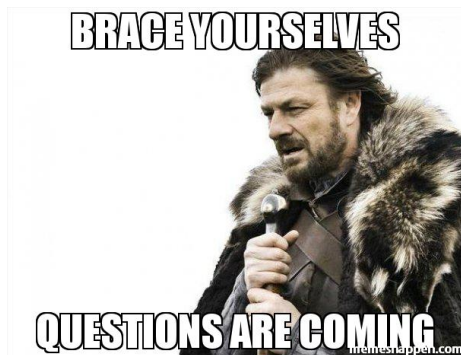




Мы знаем что мы знаем

Подбирать априорные – тяжело,
как сделать проще? Задавать себе
вопросы!

- В: Что мы знаем про
линейную регрессию?

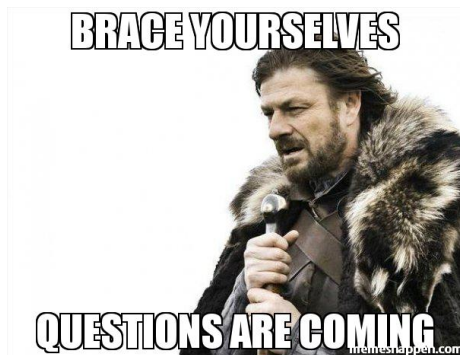




Мы знаем что мы знаем

Подбирать априорные – тяжело, как сделать проще? Задавать себе вопросы!

- В: Что мы знаем про линейную регрессию?
- О: R^2 для неё – метрика качества

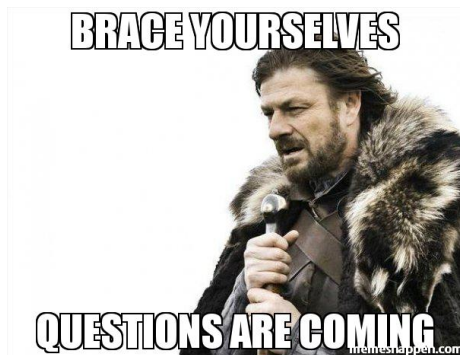




Мы знаем что мы знаем

Подбирать априорные – тяжело, как сделать проще? Задавать себе вопросы!

- В: Что мы знаем про линейную регрессию?
- О: R^2 для неё – метрика качества
- В: А еще?

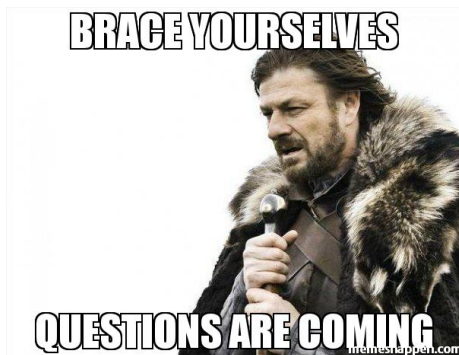




Мы знаем что мы знаем

Подбирать априорные – тяжело, как сделать проще? Задавать себе вопросы!

- В: Что мы знаем про линейную регрессию?
- О: R^2 для неё – метрика качества
- В: А еще?
- О: Какие-то переменные более важные

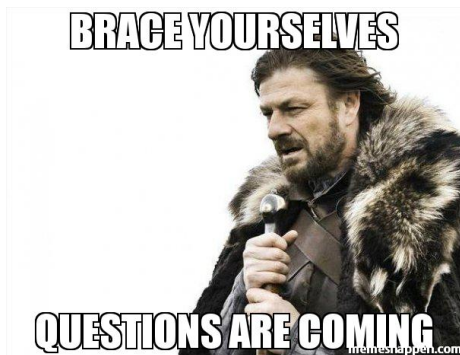




Мы знаем что мы знаем

Подбирать априорные – тяжело, как сделать проще? Задавать себе вопросы!

- В: Что мы знаем про линейную регрессию?
- О: R^2 для неё – метрика качества
- В: А еще?
- О: Какие-то переменные более важные
- О: Какие-то переменные имеют положительное влияние





Априорное для R^2

Как интерпретировать R^2 ?

- ① Мера подгонки модели под данные
 - 0 - очень плохая подгонка
 - 1 - идеальная подгонка
- ② При близких к 1 значениях – скорее всего, переобучение
- ③ R^2 - доля объясненной дисперсии

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma_T^2}$$

$$FVU = \frac{\sigma_r^2}{\sigma_T^2}$$

- σ_r^2 - сумма квадратов остатков
- σ_T^2 - общая сумма квадратов
- FVU - **F**raction **V**ariance **U**nexplained



Настраиваем априорное для R^2

Предположения для R^2 можно сделать даже перед построением регрессий, зная как они получены.

- $R^2 < 0.5$ – в поле, шум в данных
- $0.5 < R^2 < 0.75$ – в поле, хорошие данные
- $0.75 < R^2 < 0.90$ – в лаборатории, шум в данных
- $R^2 > 0.90$ – в лаборатории, хорошие данные

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.81177 -0.58567  0.05249  0.69674  2.40316

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04580    0.05694   0.804  0.42188
x1           0.42949    0.05874   7.311 2.52e-12 ***
x2           0.57386    0.06638   8.646 3.52e-16 ***
x3           0.26152    0.05773   4.530 8.58e-06 ***
x4          -0.29599    0.05444  -5.438 1.14e-07 ***
x5          -0.17564    0.05428  -3.236 0.00135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9733 on 294 degrees of freedom
Multiple R-squared:  0.4131,    Adjusted R-squared:  0.4032
F-statistic: 41.39 on 5 and 294 DF,  p-value: < 2.2e-16
```

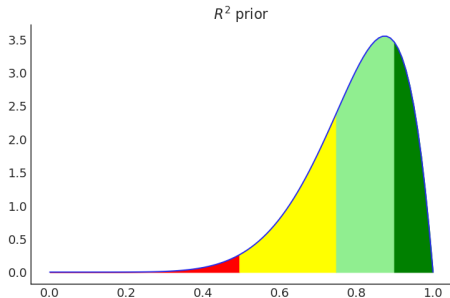




Априорный R^2

- Какого качество ваших данных?
- Вы учли все факторы для построения модели?
- Правильно ли вы собрали данные?

$$R^2 \sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)$$



Заметка

Это не **байесовский R^2** . Априорный R^2 – ожидаемый уровень качества модели.



Почувствуем разницу

Было

- Как задать априорное для β ?
- Какой у него смысл?
- Хм, нужно поменять априорное, если я добавлю параметры
- Как задать априорное для σ ?
- Слишком сложное, есть базовые значения?
- Эх, базовые значения бессмысленны

Стало

- Как хороша модель? Смотрим на R^2
- Какая переменная более или менее важная?
- Какое направление влияние будет у переменных?

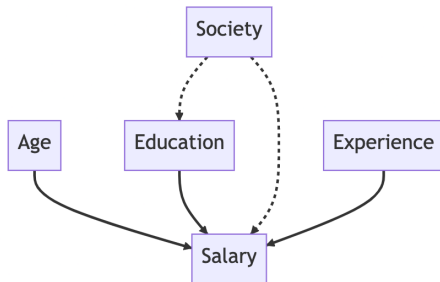


Важность переменных

Мы предсказываем зарплату,

- что важнее – возраст или опыт?
- что важнее – опыт или образование?

В традиционных моделях –
ответить можно только после
оценки





Важность переменных

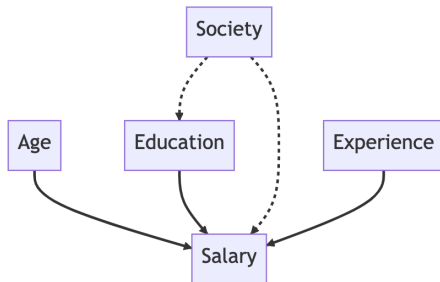
Мы предсказываем зарплату,

- что важнее – возраст или опыт?
- что важнее – опыт или образование?

В традиционных моделях – ответить можно только после оценки

Байесовский подход

- Делаем предположения, какие фичи важнее для модели
- Используем байесовские инструментальные переменные

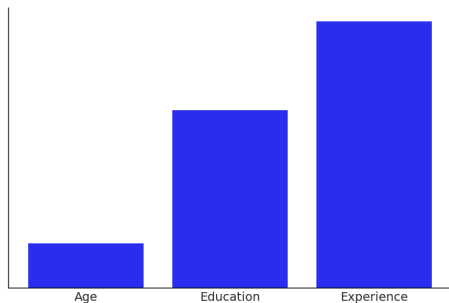




Что такое важность переменных?

Некоторые определения

- Объемы информации, которые получаем
- Fraction of **V**ariance **E**xplained





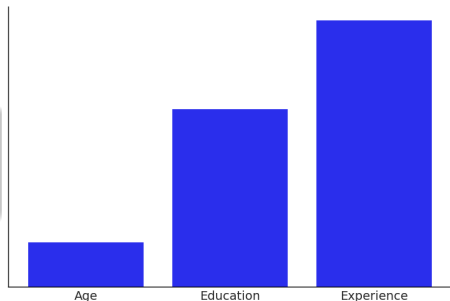
Что такое важность переменных?

Некоторые определения

- Объемы информации, которые получаем
- **Fraction of Variance Explained**

Аналогичный подход!

Как и в случае R^2 , можно считать **FVE** для фичи





Что такое важность переменных?

Некоторые определения

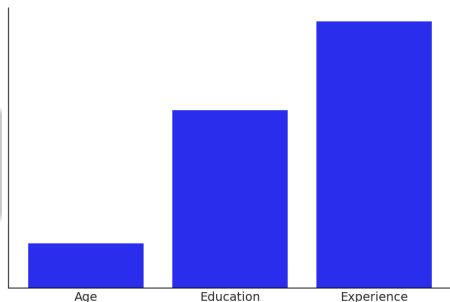
- Объемы информации, которые получаем
- Fraction of **V**ariance **E**xplained

Аналогичный подход!

Как и в случае R^2 , можно считать **FVE** для фичи

Идея проста

$$\phi_{\text{FVE}} \sim \text{Dirichlet}(\alpha_{\text{FVE}})$$



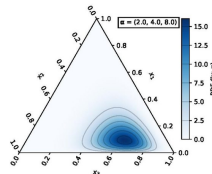
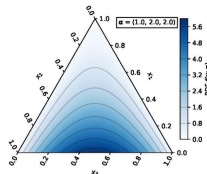
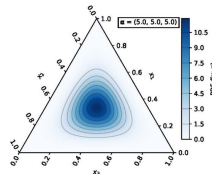
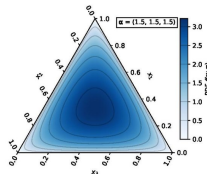
Разбираем с априорным распределением для FVE



Надо разобраться с
распределением Дирихле

$$\phi_{\text{FVE}} \sim \text{Dirichlet}(\alpha_{\text{FVE}})$$

- Чем больше α_i , тем более i -ая переменная важна
- Чем больше α_i тем больше уверенности в важности переменной





Примеры α_{FVE}

$$\phi_{FVE} \sim \text{Dirichlet}(\alpha_{FVE})$$

- $\alpha_{FVE} = (1, 1, 1)$ - Ничего не знаем про важность, возможно какие-то переменные не нужны
- $(\alpha_{FVE})_i = 1$ - переменную можно не использовать или она очень важна, не ясно
- $(\alpha_{FVE})_i = 10$ - скорее всего, переменная важна
- $(\alpha_{FVE})_i = 20$ - переменную точно надо использовать
- $\alpha_{FVE} = (10, 20, 30)$ - Используем все переменные, но 2-ая и 3-ая переменные важнее для модели



Примеры α_{FVE}

$$\phi_{FVE} \sim \text{Dirichlet}(\alpha_{FVE})$$

- $\alpha_{FVE} = (1, 1, 1)$ - Ничего не знаем про важность, возможно какие-то переменные не нужны
- $(\alpha_{FVE})_i = 1$ - переменную можно не использовать или она очень важна, не ясно
- $(\alpha_{FVE})_i = 10$ - скорее всего, переменная важна
- $(\alpha_{FVE})_i = 20$ - переменную точно надо использовать
- $\alpha_{FVE} = (10, 20, 30)$ - Используем все переменные, но 2-ая и 3-ая переменные важнее для модели

Дисклеймер

Да, это самая вольная интерпретация того, как это работает

 α_{FVE} и R^2

$$\phi_{\text{FVE}} \sim \text{Dirichlet}(\tilde{\alpha}_{\text{FVE}})$$

$$R^2 \sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)$$

Что решаем

- 1 Насколько хороша модель в принципе (R^2)
- 2 Насколько важна отдельная фича ($\tilde{\alpha}_{\text{FVE}}$)



Подводим итоги

- ① Стандартизируем данные:

$$a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$$
- ② Предположения о R^2
- ③ Предположения о важности переменных
- ④ Готово

$$\bar{y}_i \sim \mathcal{N}(\bar{\beta}^\top \bar{x}_i, \sigma)$$

$$\phi_{\text{FVE}} \sim \text{Dirichlet}(\tilde{\alpha}_{\text{FVE}})$$

$$R^2 \sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)$$

$$\sigma^2 = 1 - R^2$$

$$\bar{\beta} \sim \mathcal{N}(0, \sqrt{\phi_{\text{FVE}} \cdot R^2})$$

Больше формул

Вот недавно разработанный R2D2M2 [1], можно почитать математические выкладки.



Что-то еще можно учесть? R2D2M2CP

Да, конечно

- "Какой знак у корреляции?"
- "Насколько я уверен, что корреляция – позитивная?"



Что-то еще можно учесть? R2D2M2CP

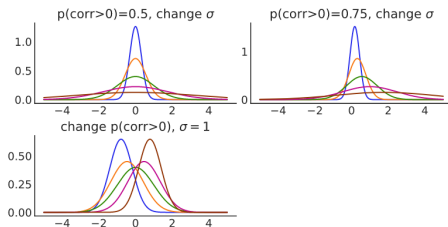
Да, конечно

- "Какой знак у корреляции?"
- "Насколько я уверен, что корреляция – позитивная?"

Какое предлагаю решение:

$$P(\bar{\beta}_j > 0) = (\psi_{CP})_j$$

$$\psi_{CP} \sim \text{Beta}(\mu = \mu_{CP}, \sigma = \sigma_{CP})$$





Технические детали

$$P(\bar{\beta}_j > 0) = (\psi_{CP})_j$$

$$\bar{\beta} \sim \mathcal{N}(\mu_{CP}(\psi_{CP}, R^2 \cdot \phi_{FVE}), \sigma_{CP}(\psi_{CP}, R^2 \cdot \phi_{FVE}))$$

$$\psi_{CP} \sim \text{Beta}(\mu = \mu_{CP}, \sigma = \sigma_{CP})$$

$$\phi_{FVE} \sim \text{Dirichlet}(\tilde{\alpha}_{FVE})$$

$$R^2 \sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)$$

Значения μ_{CP} , σ_{CP} – единственные

$$\begin{cases} \mu_{CP}(p, v) = \frac{\sqrt{2v} \operatorname{erf}^{-1}(2p-1)}{\sqrt{2 \operatorname{erf}^{-1}(2p-1)^2 + 1}} \\ \sigma_{CP}(p, v) = \frac{\sqrt{v}}{\sqrt{2 \operatorname{erf}^{-1}(2p-1)^2 + 1}} \end{cases}$$



Подводим итоги

Чтобы R2D2M2CP работал

- 1 Стандартизируем данные:
$$a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$$
- 2 Предположение о R^2
- 3 Предположение о важности переменных
- 4 Предположение о знаке корреляций между переменным
- 5 Точно готово!

Пример практического применения
[2]



<https://github.com/pymc-devs/pymc-experimental/pull/137>



Вернем к обобщенным линейным моделям

Рассмотрим следующие предпосылки:

$$y_i \sim \mathcal{T}(\nu_i, m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i + \dots$$



Вернем к обобщенным линейным моделям

Рассмотрим следующие предпосылки:

$$y_i \sim \mathcal{T}(\nu_i, m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i + \dots$$

- Какие переменные важны s ?
(предположение о важности переменных)

Априорное для N_i

Число степеней свободы можно рассчитать при следующем предположении:

<https://github.com/pymc-devs/pymc-experimental/pull/252>



Вернем к обобщенным линейным моделям

Рассмотрим следующие предпосылки:

$$y_i \sim \mathcal{T}(\nu_i, m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i + \dots$$

- Какие переменные важны s ?
(предположение о важности переменных)
- Они вообще как-то влияют? (предположение о R^2)

Априорное для N_i

Число степеней свободы можно рассчитать при следующем предположении:

<https://github.com/pymc-devs/pymc-experimental/pull/252>



Финальные заметки

- R2D2M2CP – труднопроизносимый термин
- Дает о чем подумать при использовании традиционных моделей
- Расширяет применение обобщенных линейных моделей, позволяя контролировать вспомогательные модели
- Дополнительные инструмент для работы с обобщенными аддитивными моделями с применением гауссовских процессов

LM

$$\hat{\beta} = (X^T X)^{-1} X^T y$$



GLM

$$y_i \sim \mathcal{N}(m_i, s_i)$$

$$m_i \sim x_i + \dots$$

$$\log s_i \sim z_i$$



R2D2M2CP

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma_T^2}$$

$$FVU = \frac{\sigma_r^2}{\sigma_T^2}$$



GLM

+

R2D2M2CP





Ссылки I



J. E. Aguilar and P.-C. Bürkner.

Intuitive joint priors for bayesian linear multilevel models: The r2d2m2 prior, 2023.



M. Kochurov.

pymc-devs/pymc-experimental: Pull Request 137 R2D2M2CP.
GitHub, 2023.