Эконометрика: Гаусс-Марков

На этой неделе мы пытаемся понять, зачем нам в 2023 году линейная регрессия, какие у неё есть возможности, а также какие проблемы есть в классической модели линейной регрессии.

6.1 Модель линейной регрессии

Говорят, что слово "регрессия" изначально возникало в работах Френсиса Гальтона в контексте его работ, связанных с исследованием зависимости роста сыновей от роста отцов. Когда мы смотрим на годы жизни Гальтона, становится ясно, насколько эта идея стара. Тем не менее, даже в 2023 году линейная регрессия не только стала классикой, но и остаётся основным, если не главным, инструментом аналитики.

В классической модели полагается, что есть некоторая вещественнозначная случайная величина Y, которая зависит от другой векторной случайной величины X, принимающей значения в $\mathbb{R}^{1\times k}$, как

$$Y = X\theta + \varepsilon,$$

где ε – это случайная величина (шум), про которую в самом общем случае предполагается, что она обладает нулевым средним и конечной дисперсией.

Для удобства исследования часто полагают, что предоставлен набор данных, состоящий из n реализаций (X_i, Y_i). Поскольку исследователя не интересует как устроена случайная величина X (действительно, зачем нам для анализа прибыли от продажи компьютеров знать *как именно* клиент получил тот доход, который он указывал в анкете), её при анализе регрессии предполагают *фиксированной константой*; таким образом, случайность приходит только из шумов ε_i. В матричном виде обычно модель линейной регрессии записывается как

$$Y = X\theta + \varepsilon,$$

где $X \in \mathbb{R}^{n\times k}$ – матрица наблюдений (ещё её называют матрицей плана), ε – случайный вектор шумов, который принимает значения в $\mathbb{R}^n$, а $\theta \in \mathbb{R}^k$ – неизвестный параметр модели. Иногда отдельно выписывают константу, но на самом деле её можно просто вписать в виде столбца единиц в матрицу X.

Итак, мы получили вероятностную модель, данные, следовательно, мы можем начать задумываться о том, как оценить θ. Первая мысль, которая возникает при взгляде на уравнение выше такая: случайность только в ε. Вторая мысль – если бы мы знали настоящий θ, мы могли бы вычислить $\varepsilon = Y - X\theta$. Но третья состоит в том, что неизвестный θ нисколько не мешает выписать функцию лог-правдоподобия:

$$l(\theta) = \ln p_\varepsilon(Y - X\theta),$$

где p_ε – совместное распределение (плотность) шумов. Таким образом, нам неизбежно нужно про него что-то предположить, если мы хотим получить алгоритм оценки.

Теорема 6.1. *Пусть* $\varepsilon_i \sim^{iid} \mathcal{N}(0, 1)$, *тогда оценка метода максимального правдоподобия*

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

является несмещённой.

▷ Здесь нужно вспомнить, что поскольку распределение шумов гауссовское с единичной ковариационной матрицей, попытка прооптимизировать лог-правдоподобие – это то же самое, что записать

$$\hat{\theta} = \arg\min_\theta \|Y - X\theta\|_2^2,$$

то есть, сделать оценку *методом наименьших квадратов*. Для этого нужно взять градиент, приравнять нулю и получить

$$X^T(X\theta - Y) = 0,$$

откуда следует результат. Для проверки несмещённости возьмём матожидание (помним, что случайность только в ε):

$$\mathbb{E}\left[(X^T X)^{-1} X^T Y\right] = \mathbb{E}\left[(X^T X)^{-1} X^T (X\theta + \varepsilon)\right] = \theta.$$

□

Заметим, что если отказаться от единичной дисперсии в пользу, например σ^2, то и в таком случае мы можем записать метод максимального правдоподобия и выписать ещё дисперсию оценки

$$\mathrm{Var}\left[\hat{\theta}\right] = \sigma^2 (X^T X)^{-1}.$$

6.2 Теорема Гаусса-Маркова

Когда мы говорим о классических предпосылках линейной регрессии, имеются в виду предположения теоремы Гаусса-Маркова:

1. Модель корректна: все признаки, влияющие на Y учтены;
2. X детерминирована и колонки линейно независимы;
3. ε_i независимы в совокупности и одинаково распределены, не зависят от наблюдений X_i, имеют нулевое матожидание и одинаковую дисперсию σ^2.

Если предположения верны, то выполнена

Теорема 6.2. *(Гаусс-Марков) При верных предположениях оценка методом наименьших квадратов является несмещённой и эффективной (в классе всех несмещённых оценок, Best Linear Unbiased Estimate).*

Верны ли эти предположения в жизни? На самом деле, не очень и этому есть практические свидетельства. Тем не менее, давайте повременим с тем, чтобы отметать линейную регрессию как анахронизм и для начала посмотрим, какими свойствами она обладает и стоит ли пытаться её спасать.

6.3 Проверка гипотез в классической линейной регрессии

Первое и одно из главных преимуществ классической модели состоит в том, что там достаточно простой аппарат проверки статистических гипотез о параметрах.

6.3.1 Гипотеза об одном параметре

Что, например, даёт гауссовость шума? Если $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, то

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2 (X^T X)^{-1}),$$

что позволяет уже думать о статистических критериях. Конечно, никто нам не даёт настоящую σ, поэтому приходится пользоваться оценкой

$$\sigma^2 (X^T X)^{-1} \approx \hat{\sigma}^2 (X^T X)^{-1}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i))^2.$$

Здесь мы впервые ввели оценку для дисперсии $\hat{\sigma}^2$, которая оценивается через средний квадрат остатков регрессии. Оказывается, что если истинный параметр равен θ_0, то в силу несмещённости оценки и предполагаемой гауссовости остатков статистика

$$T = \frac{\hat{\theta}_j - \theta_j^0}{\hat{\sigma}(\hat{\theta}_j)} \sim t(n - k - 1),$$

где k – число параметров, а $\hat{\sigma}(\hat{\theta}_j)^2$ – это j-й элемент на диагонали матрицы $\hat{\sigma}^2 (X^T X)^{-1}$. Такая статистика используется для проверки гипотезы

$$H_0 : \theta_j = \theta_j^0, H_A : \theta_j \neq \theta_j^0$$

построенный критерий для проверки гипотезы об одном параметре называется тестом Стьюдента или t-тестом, который уже встречался нам ранее. Если же шум не гауссовский, то свойства метода максимального правдоподобия нам гарантируют, что асимптотически при больших n статистика $T \sim \mathcal{N}(0,1)$ и мы получаем асимптотический z-тест.

Пример 6.1. *Положим, мы хотим по числовому набору данных о росте спортсменов и их результатах прыжков в длину понять, как связаны рост и результат прыжка. Если для набора данных выполнены предположения регрессии (на следующей неделе мы обсудим, как их можно проверять), то мы можем, оценив регрессию*

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

проверить с помощью t- или z-теста гипотезу

$$H_0 : \beta = 0, \quad H_A : \beta \neq 0.$$

Если возобладает альтернатива, то делаем вывод, что есть линейная зависимость и она значима. На самом деле, что-то похожее на выборочный коэффициент корреляции

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

является оценкой параметра β.

6.3.2 Гипотеза о многих параметрах

Естественный вопрос состоит в том, можно ли проверять одну гипотезу про несколько параметров сразу. Здесь есть идея оценить две модели линейной регрессии: одну при фиксированном наборе параметров(верной гипотезе) и одну при полном наборе параметров, – и сравнить их как-то между собой. Например, можно сопоставить ошибки прогнозов. Для этого вводятся

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \; TSS = \sum_{i=1}^n (y_i - \overline{y})^2, \; ESS = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2,$$

которые соответственно более развернуто называются *residual, total, explained sum of squares*. RSS – это суммарная квадратичная ошибка прогноза полной оценённой модели регрессии, TSS – суммарная ошибка тривиального бэйзлайна (просто выборочное среднее Y, посчитанное по данным), ESS – различие прогнозов полной модели и тривиального бэйзлайна (отсюда в названии *explained*). Коэффициент детерминации

$$R^2 = \frac{ESS}{TSS}$$

вводится как доля *объяснённой* более сложной моделью ошибки в ошибке тривиального бейзлайна. Он лежит в отрезке $[0,1]$, так как предполагается, что в более сложную модель

уже включена константа, которая как раз оценивается $\overline{y}$. Использовать R^2 для оценки качества регрессии нужно с осторожностью, потому что он отражает скорее не качество, а полноту: при добавлении новых параметров в модель R^2 будет только расти.

Чем больше R^2, тем полнее модель; на этом можно сделать статистический тест на общую значимость оценённой регрессионной модели, то есть, для проверки гипотезы

$$H_0 : \theta = 0, \quad H_A : \theta \neq 0,$$

иными словами: "Мы вообще что-то адекватное оцениваем или модель надо полностью поменять?" Если принимаем гипотезу, то модель вообще ни на что не способна, а если выбираем альтернативу, то что-то из параметров значимое там всё-таки есть, а конкретнее про каждый можно выяснить с помощью t- или z-теста, как ранее. Мы в результате в гауссовской модели получаем точный критерий, называемый критерием Фишера или F-тестом, потому что статистика

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

при верной гипотезе имеет распределение Фишера $F(k-1, n-k)$ с k, равным количеству параметров в линейной регрессии, n – размер выборки. В негауссовском случае статистика $(k-1)F$ имеет асимптотическое при больших n распределение $\chi^2(k-1)$.

И это ещё не всё. По такой же логике мы можем сравнивать две разных модели линейной регрессии: более сложную и более простую, которая по представленности признаков содержится в более сложной. Об этом подробнее на семинаре.

6.4 Зачем нужна линейная регрессия в 2023?

Можно спросить: действительно, зачем? Предсказательно линейные модели часто слабые и уступают, например, алгоритмам бустинга или нейросетям, это не говоря про Deep Learning, который в принципе решает структурно сложные задачи, которые линейной модели недоступны. Предсказательная сила линейной модели действительно не всегда самая сильная, хотя её можно улучшать, подбирая правильно признаки. Но линейная модель обладает очень важным для многих свойством: интерпретируемостью. Более того, мы имеем арсенал статистических критериев, которые позволяют использовать достаточно сильные аргументы для обоснования адекватности полученной модели. Представьте себе, как отвечать на вопрос: "Почему здесь ваша сложная нейросеть предсказала, что нам нужно вложить 5 миллионов долларов?"

В анализе данных, где участвуют различные вероятностные оцениваемые модели, можно ставить фундаментально две разных задачи.

1. Задача предсказания (*что будет завтра?*): требуется построить модель, которая будет давать как можно лучший прогноз на новых данных; основной метрикой является обобщающая способность модели, отсюда в машинном обучении идея разделения выборки на train-test и кросс-валидация.

2. Задача описания данных (*как устроен мир?*): требуется построить модель с понятной взаимосвязью различных факторов и которая могла бы помочь ответить на вопрос о том, что будет, если один или несколько факторов изменятся, и результаты которой можно было бы обсуждать в контексте существующей науки или общей практики.

Машинное обучение почти целиком про первую задачу, регрессия во многом про вторую. Тем не менее, даже в задаче предсказаний (особенно это актуально в экономических временных рядах) линейные регрессионные модели (типа ARiMA, например), с одной стороны, являются очень простым решением (легко оценить и использовать как бэйзлайн), а с другой – в некоторых случаях не бьются классическими методами машинного обучения из-за необычной по масштабу структуры времени.

Как ни посмотреть, модель линейной регрессии обладает многими преимуществами, которых нет у других предсказательных аглоритмов, поэтому ей важно и нужно заниматься, обходя и ремонтируя проблемы в предположениях классической модели. Этим и занимается эконометрика.