КЛАССИЧЕСКИЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Оценки модели линейной регрессии

Построение модели

Слайд 2

Пусть n-вектор $\mathbf{Y}$, связан с q-мерной *неслучайной векторной переменной* $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_q]'$. Значения Y_i, $i = 1, \ldots, n$, полученные в эксперименте при заданных $\mathbf{x}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iq}]'$, случайным образом изменяются около некоторого неизвестного истинного уровня $\eta(\mathbf{x}_i)$. Тогда можем записать

$$Y_i = \eta(\mathbf{x}_i) + \varepsilon_i,\ i = 1, \ldots, n, \qquad (1)$$

где ε_i – случайная ошибка, которая объясняет отклонение Y_i от величины $\eta(\mathbf{x}_i)$. При этом $\boldsymbol{\varepsilon}$ может быть случайной компонентой, присущей величине $\eta(\mathbf{x})$, и представлять случайную ошибку измерения значений $\mathbf{Y}$ или влияние различных неучтенных факторов. Предположим, что $\eta(\mathbf{x})$ можно описать линейной моделью первого порядка по $\mathbf{x}_j$ с q переменными

$$\eta(\mathbf{x},\boldsymbol{\beta}) = \beta_0 + \beta_1\mathbf{x}_1 + \ldots + \beta_q\mathbf{x}_q.$$

где $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_q]'$ – вектор неизвестных параметров (коэффициентов), подлежащий оцениванию. Тогда получим

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{x}_1 + \ldots + \beta_q\mathbf{x}_q + \boldsymbol{\varepsilon} \qquad (2)$$

Если в формуле (2) $\mathbf{M}\boldsymbol{\varepsilon} = 0$, то условное математическое ожидание случайного вектора $\mathbf{Y}$ при заданных переменных $\mathbf{x}_j$, $j = 1, \ldots, q$ равно

$$\mathbf{M}[\mathbf{Y}/\mathbf{x}] = \beta_0 + \beta_1\mathbf{x}_1 + \ldots + \beta_q\mathbf{x}_q \qquad (3)$$

Уравнение (3), в котором $\mathbf{x}$ играет роль «независимой» переменной, называется *уравнением регрессии* или просто *регрессией*.

Термин «регрессия» впервые был введен Ф. Гальтоном[1] (1886) в теории наследственности для обозначения явления «*возврата к среднему состоянию*» (*regression to mediocrity*), состоящего в том, что дети тех родителей, рост которых превышает среднее значение на a единиц, имеют в среднем рост, превышающий среднее значение меньше чем на a единиц.

В дальнейшем переменные[1] $\mathbf{Y}$, $\mathbf{x}_j$, $j = 1, \ldots, q$ и $\boldsymbol{\varepsilon}$ будем называть *откликом*, *регрессорами* и *остатком* (используются другие названия этих переменных: выход, зависимая или эндогенная переменная; факторы, предикторы, входные, экзогенные или независимые переменные; ошибка, помеха, невязка).

Когда используется уравнение (2) при анализе совокупности данных и оценивается вектор параметров $\boldsymbol{\beta}$, то предполагается, что элементы этой совокупности *однородны* в смысле подчинения одному и тому же причинному закону. Это означает, что параметры β_j приемлемы для каждого отдельно взятого наблюдения.

Слайд 3

***Пример* 1.** Идентификация статических характеристик сложного объекта, выходы которого, измеряемые со случайными ошибками, является функциями многих входных переменных.

---

[1] Гальтон (Galton) Фрэнсис (16.02.1822 – 17.01.1911) – английский психолог и антрополог.

[1] Следуя большинству книг по регрессионному анализу, случайный вектор-отклик и матрицы будем обозначать полужирными прописными буквами; векторы-регрессоры и вектор-остаток строчными полужирными буквами.

Необходимо по наблюдениям входов и выходов определить эти функции. В общем случае совокупность переменных, определяющих текущее состояние сложного объекта, можно описать следующими группами входных и выходных переменных (рис. 1).

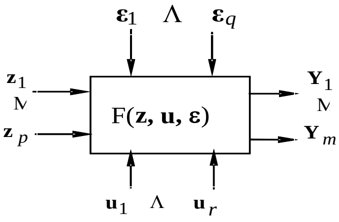

Рис. 1. Модель сложного объекта

1. *Контролируемые неуправляемые* переменные $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p)$, значения этих переменных можно измерить, но нельзя произвольно изменить.

2. *Контролируемые управляющие переменные* $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$, значения, которых в любой момент времени можно изменить в пределах допустимого диапазона.

3. *Неконтролируемые неуправляемые переменные* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_q)$, которые характеризуют множество реально существующих факторов, влияющих на текущее состояние объекта, но недоступных контролю и управлению.

4. *Контролируемые управляемые переменные* $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$, которые характеризуют результат функционирования объекта.

Входные переменные **z, u, ε** могут рассматриваться как причины, оказывающие влияние на каждую из выходных переменных $\mathbf{Y}_i$.

При общем рассмотрении нет необходимости разделять контролируемые переменные (**z, u,**) поэтому объединим их в одну группу и обозначим **X**. Далее будем полагать, что $\mathbf{x}_j$ при $j = 1, \ldots, q$ – неслучайные контролируемые независимые переменные; **ε** – случайная неконтролируемая переменная (остаток, помеха, ошибка). Так как каждая из выходных переменных $\mathbf{Y}_i$ полностью определяется в вероятностном смысле группой входных переменных **X** и остатком **ε,** то достаточно рассмотреть схему с одной выходной переменной (откликом). Будем полагать, что случайная остаток **ε** аддитивно приложен к выходной переменной **Y**, т. е. $\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$. Тогда физическую модель, характеризующую зависимость **Y** от **X** можно выразить уравнением (1).

Слайд 4

Структурная схема объекта, соответствующая этой модели, приведена на рис. 2.

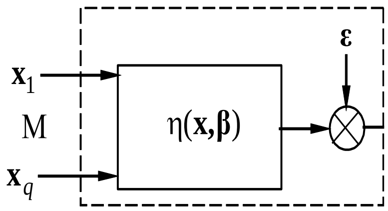

Рис. 2. Структурная схема объекта

В общем случае функция **η(x,β)** нелинейна относительно вектора параметров **β**. Простейшим и важнейшим для практики является случай линейной зависимости **η(x,β)** от **β**. Линейную регрессионную модель можно получить, разложив **η(x,β)** в ряд Тейлора в точке $\mathbf{x}_0 = 0$.

Тогда

$$Y_i = \eta(0,\mathrm{K}\,,0) + \sum_{j=1}^{q}\left(\frac{\partial\eta}{\partial x_j}\right)_0 x_{ij} + \frac{1}{2}\sum_j\sum_k\left(\frac{\partial^2\eta}{\partial x_j \partial x_k}\right)_0 x_{ij}x_{ik} + \mathrm{K} + \varepsilon_i\,.$$

Обозначив постоянные

$$\beta_0 = \ (0,\mathrm{K}\,,0),\quad \beta_j = (\partial\eta/\partial x_j)_0,\quad \beta_{jk} = (1/2)(\partial^2\eta/\partial x_j \partial x_k)_0\,,$$

получим

$$Y_i = \beta_0 + \sum\nolimits_{j=1}^{q}\beta_j x_{ij} + \sum\nolimits_j\sum\nolimits_k \beta_{jk}x_{ij}x_{ik} + \mathrm{K} + \varepsilon_i\,.$$

Ограничимся рассмотрением в этом уравнении только первых двух членов, случайные ошибки и ошибки за счет неучтенных членов ряда отнесем к остатку $\boldsymbol{\varepsilon}$. При этом будем полагать, что неучтенные члены *не коррелированы* с учтенными. Тогда уравнение можно переписать в виде модели линейной регрессии (2). ■

Модель вида (2) является весьма общей и очень широко используется. Частными случаями ее являются, например, полиномиальная модель q-го порядка одной переменной

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_q x_i^q + \varepsilon_i,$$

Основное свойство модели вида (2) заключается в ее *линейности* по отношению к неизвестному вектору коэффициентов $\boldsymbol{\beta}$. По сравнению с ней, например, модель

$$Y_i = \beta_0 + \beta_1 e^{-\beta_2 x_i} + \varepsilon_i$$

нелинейная по параметру β_2.

Слайд 5

Рассмотрим оценки $\hat{\boldsymbol{\beta}}$ вектора коэффициентов β регрессионной модели (3). При этом будем различать два типа оценок. Первый – точечные оценки, получаемые на основании наблюдаемых данных регрессоров и отклика. Второй тип оценок связан с построением доверительных областей (интервалов) в пространстве оценок, которые с заданной вероятностью «накрывают» неизвестное истинное значение. Анализ уравнения (3) и оценку его коэффициентов будем проводить с использованием матричной алгебры. Применение матриц упрощает расчеты и придает им наглядность.

Оценивание параметров. Свойства оценок

Рассмотрим схему, изображенную на рис. 2, где пунктиром выделена ненаблюдаемая часть. Пусть отклик $\mathbf{Y}$ связан с входами полиномом вида (2). Записывая эти n уравнений в матричной форме, получаем

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1q} \\ x_{20} & x_{21} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nq} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Или

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

где $x_{10} = x_{20} = \ldots = x_{n0} = 1$. Матрица $\mathbf{X}$ типа «объект-признак» (см. п. 1.1) размера $n\times(q + 1)$ называется *регрессионной матрицей*, столбцами которой являются регрессоры $\mathbf{x}_j$, $j = 1, \ldots, q$, а строками – n объектов или опытов; $\mathbf{Y}$ и $\boldsymbol{\varepsilon}$ – n-векторы отклика и остатков, $\boldsymbol{\beta}$ – подлежащий оцениванию $(q + 1)$-вектор неизвестных коэффициентов. В активных экспериментах элементы матрицы $\mathbf{X}$ выбираются равными только нулю и единице и в этом случае $\mathbf{X}$ называется *матрицей плана*.

Необходимо по наблюдениям $(x_{i1}, \ldots, x_{iq}, Y_i)$, $i = 1, \ldots, n$ найти наилучшую оценку $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q)'$ вектора коэффициентов $\boldsymbol{\beta}$ = $\beta_0, \beta_1, \ldots, \beta_q)'$ уравнения регрессии $\boldsymbol{\eta}(\mathbf{x},\boldsymbol{\beta})$. Одним из самых распространенных методов оценки вектора коэффициентов регрессии $\boldsymbol{\beta}$ является *метод наименьших квадратов* (МНК) (*least squares method*). Для обеспечения эффективности МНК-оценок должны соблюдаться следующие постулаты[1]:

---

[1] В реальной жизни все эти постулаты редко соблюдаются.

Слайд 6

1. Число регрессоров q полинома (3) априори известно точно.
2. Все регрессоры измеряются без ошибок, а вычисления проводятся абсолютно точно.
3. Остаток $\boldsymbol{\varepsilon}$ является независимой нормально распределенной случайной величиной с нулевым средним $\mathbf{M}[\boldsymbol{\varepsilon}] = 0$ и неизвестной постоянной дисперсией σ_ε^2 при всех $i = 1, \ldots, n$.
4. Дисперсия отклика Y_i постоянна, или является известной функцией номера наблюдения $i = 1, \ldots, n$.
5. Распределение Y_i одинаково при всех $i = 1, \ldots, n$.
6. Число опытов n существенно больше числа регрессоров q.

Слайд 7

Если постулаты (1 – 6) соблюдаются и $\mathbf{X'X}$ обратимая матрица, то согласно фундаментальной *теореме Гаусса-Маркова* наилучшей оценкой вектора коэффициентов $\boldsymbol{\beta}$ является оценка $\hat{\boldsymbol{\beta}}$, доставляющая *минимум суммы квадратов остатков* (невязок, ошибок, помех):

$$Q(\hat{\beta}) = \sum_{i=1}^{n}\left(Y_i - \sum_{k=0}^{q} x_{ik}\hat{\beta}_k\right)^2 \to \min .$$

Заметим, что этот остаток не может равняться нулю, так как число наблюдений n превосходит число неизвестных параметров q. Если $Q(\hat{\beta})$ имеет производные по $\hat{\beta}_m$, то необходимым условием минимума являются уравнения

$$\partial Q / \partial\hat{\beta}_m = -2\sum_{i=1}^{n}\left(Y_i - \sum_{k=0}^{q} x_{ik}\hat{\beta}_k\right) x_{im} = 0 ,$$

или

$$\sum_{i=1}^{n} x_{im} \sum_{k=0}^{q} x_{ik}\hat{\beta}_k = \sum_{i=1}^{n} x_{im} Y_i ,\; m = 0, 1, \ldots, q . \tag{5}$$

Система уравнений (5) называется *системой нормальных уравнений* (СНУ) МНК. Слово «нормальных» не связано с нормальным распределением вероятностей, а только подчеркивает, что уравнения, как правило, имеют такой «нормальный» вид.

Слайд 8

Обозначим $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Будем минимизировать величину $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \|\mathbf{Y} - \boldsymbol{\theta}\|^2$ по отношению к $\boldsymbol{\theta} \in \Omega$, где Ω – подпространство оценок $\hat{\boldsymbol{\theta}}$. Если изменять значения вектора $\boldsymbol{\theta}$ в пределах Ω, то квадрат длины вектора $\|\mathbf{Y} - \boldsymbol{\theta}\|^2$ достигнет минимума при значении $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, которое является проекцией вектора $\mathbf{Y}$ на подпространство Ω. Тогда справедливо $(\mathbf{Y} - \hat{\boldsymbol{\theta}}) \perp \hat{\boldsymbol{\theta}}$ и, следовательно, $(\mathbf{Y} - \hat{\boldsymbol{\theta}}) \perp \mathbf{X}$ (рис. 3). Отсюда для скалярного произведения $(\mathbf{Y} - \hat{\boldsymbol{\theta}})$ и $\mathbf{X}$ получаем

$$(\mathbf{Y} - \hat{\boldsymbol{\theta}}, \mathbf{X}) = \mathbf{0}$$

Или

$$\mathbf{X}'\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{Y} . \tag{6}$$

Если столбцы матрицы **X** линейно независимы, то существует единственный вектор параметров $\boldsymbol{\beta}$, для которого $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Тогда система (6) выразится в виде СНУ $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$.

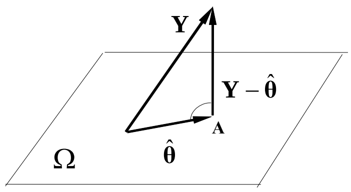

Рис. 3. Нахождение точки А, для которой норма $\|\mathbf{Y} - \hat{\boldsymbol{\theta}}\|$ минимальна

Слайд 9

Обозначим $\mathbf{C} = \mathbf{X}'\mathbf{X}$, , $\boldsymbol{\Psi} = \mathbf{X}'\mathbf{Y}$, тогда СНУ запишется в виде

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\Psi}. \tag{7}$$

Свойства матрицы **C**:

a) так как регрессоры $\mathbf{x}_1, \ldots, \mathbf{x}_q$ линейно независимы, то матрица **C** невырождена;

b) **C** положительно определена и ранг ее в точности равен q;

c) **C**–симметричная матрица, $\mathbf{C} = \mathbf{C}'$, т. е. является э*рмитовой*[1].

Отсюда следует, что СНУ (6) имеет единственное решение

$$\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\boldsymbol{\Psi} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{8}$$

Оценка $\hat{\boldsymbol{\beta}}$ называется *оценкой метода наименьших квадратов* (МНК-о*ценкой*). Так как в решении (8) матрица $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ неслучайная, то $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y}$ является линейной комбинацией наблюдений **Y**. В соответствии с теоремой Гаусса-Маркова МНК-оценка $\hat{\boldsymbol{\beta}}$ имеет наименьшую дисперсию среди всех возможных несмещенных линейных оценок.

Слайд 10

Свойства оценок

МНК-оценка $\hat{\boldsymbol{\beta}}$ является случайной величиной. Найдем математическое ожидание оценки $\hat{\boldsymbol{\beta}}$. Используя решение (8) и, учитывая, что матрица **X** является детерминированной, получаем

$$\mathbf{M}[\hat{\boldsymbol{\beta}}] = \mathbf{M}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \mathbf{M}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] =$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}.$$

Таким образом, математическое ожидание оценки вектора $\hat{\boldsymbol{\beta}}$ равно истинному значению $\boldsymbol{\beta}$, т. е. $\hat{\boldsymbol{\beta}}$ является *несмещенной оценкой.* Другими словами, если эксперимент снова и снова повторяется при неизменной матрице **X**, среднее значение $\hat{\boldsymbol{\beta}}$ будет равно $\boldsymbol{\beta}$.

Распределения

До сих пор единственное предположение относительно $\boldsymbol{\varepsilon}$ состояло в том, что $\mathbf{M}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $D[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2\mathbf{I}_n$. Если дополнительно предположить, что остаток $\boldsymbol{\varepsilon}$ нормально распределен с параметрами 0, $\sigma_\varepsilon^2\mathbf{I}_n$, или, при кратком обозначении, $\boldsymbol{\varepsilon} : N_n(0, \sigma_\varepsilon^2\mathbf{I}_n)$, то $\mathbf{Y} : N_n(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2\mathbf{I}_n)$. Отсюда получается целый ряд результатов,
связанных с распределениями. Если $\mathbf{Y} : N_n(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2\mathbf{I}_n)$, то:

1) $\hat{\boldsymbol{\beta}} \sim N_n(\boldsymbol{\beta}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1})$,

2) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma_\varepsilon^2 \sim \chi^2_{(q+1)}$,

3) $\hat{\boldsymbol{\beta}}$ не зависит от S_e^2,

4) $SSE/\sigma_\varepsilon^2 = (n-q-1)S_e^2/\sigma_\varepsilon^2 \sim \chi^2_{(n-q-1)}$.

[1] Эрмит (Hermite) Шарль (24.12.1822 – 14.01.1901) – французский математик.

Предположение нормальности распределения остатков позволяет создать целостную систему статистической обработки, которая включает точечные, интервальные оценки и проверки статистических гипотез. Однако на практике распространенный миф нормальности распределения не всегда выполняется, а в случае малых выборок гипотезу нормальности распределения ошибок трудно проверить. Отклонение от нормальности может быть вызвано и засорением наблюдений чужеродными элементами. В этом случае для обнаружения и удаления этих элементов нужно применить методы, изложенные в предыдущей лекции.

Другой подход связан с применением вместо МНК *метода наименьших модулей* (МНМ). Близким к МНМ является *непараметрический регрессионный анализ*, например, *знаковый регрессионный анализ*, который позволяет получать хорошие оценки и при сильно засоренных выборках. И, наконец, для таких данных можно использовать *робастную регрессию* или решать задачу регрессии с помощью *нейронных сетей*.