Лабораторная работа № 3. Модель линейной регрессии

К регрессионному анализу относятся задачи выявления искаженной "шумом" функциональной зависимости интересующего исследователя показателя Y от измеряемых переменных $X_1, X_2, \ldots, X_m$. Данными служит таблица экспериментально полученных "зашумленных" значений Y на разных наборах $X_1, X_2, \ldots, X_m$. Основной целью обычно является как можно более точный прогноз (предсказание) Y на основе измеряемых (предикторных) переменных.

Под линейной регрессией понимают ситуацию, когда зависимость Y (отклика) от предикторных переменных $X_1, X_2, \ldots, X_m$ линейная, но наблюдается Y со случайной ошибкой ("шумом").

Y — целевая переменная (отклик), $X_1, X_2, \ldots, X_m$ — объясняющие переменные (факторы).

Модель:

$$Y_i = X_{i,1}\theta_1 + X_{i,2}\theta_2 + \cdots + X_{i,n}\theta_m + \varepsilon_i, \quad i = 1 \ldots n, n \geq m, \tag{1}$$

$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ – вектор случайных ошибок ("шум").

Ввдем обозначения

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ — вектор наблюдений

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$ — вектор неизвестных значений параметров

$\mathbb{X} = \|X_{i,j}\|, \ i = 1, \ldots, n, \ j = 1, \ldots, m$ — матрица плана.

Тогда (1) можно записать в векторной форме:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \tag{2}$$

Предположения:

$(A.1)$ Столбцы $X_{(j)} = (X_{1,j}, \ldots, X_{n,j})^T$, $j = 1, \ldots, m$, матрицы плана $\mathbb{X}$ линейно независимы. Иными словами, ввиду выполнения неравенства $n \geq m$ матрица $\mathbb{X}$ имеет ранг m.

$(A.2)$ Случайные величины $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ одинаково распределены с $\mathbf{E}\varepsilon_i = 0, \ \mathbf{D}\varepsilon_i = \sigma^2$ (параметр $0 < \sigma < \infty$ также неизвестен) и *некоррелированы:* $\mathbf{E}(\varepsilon_i\varepsilon_j) = 0$ при $j \neq j$.

Оценим параметры $\theta_1, \theta_2, \ldots, \theta_m$ **методом наименьших квадратов**, минимизируя по $\boldsymbol{\theta}$ функцию

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{n}(Y_i - \theta_1 X_{i,1} - \cdots - \theta_m X_{i,m})^2 = (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\theta})^T(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\theta}). \tag{3}$$

Точка ее минимума $\widehat{\boldsymbol{\theta}}$ называется *МНК-оценкой*, вектор $\widehat{\boldsymbol{\delta}}$ с $\delta_i = Y_i - \widehat{\theta}_1 X_{i,1} - \cdots - \widehat{\theta}_m X_{i,m}$, $i = 1, \ldots, n$ -вектором остатков, в векторной форме $\widehat{\boldsymbol{\delta}} = \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}}$.

Величина $RSS = F(\widehat{\boldsymbol{\theta}}) = (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}})^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\delta}}$ называется *остаточной суммой квадратов.*

Положительно определенная матрица $\mathbf{B} = \mathbb{X}^T\mathbb{X}$, называемая *информационной*, является (ввиду предположения $(A.1)$) невырожденной. Из геометрических соображений легко находим, что оценка $\widehat{\boldsymbol{\theta}}$, минимизирующая значение $F(\boldsymbol{\theta})$ определяет проекцию $\mathbb{X}\widehat{\boldsymbol{\theta}}$ вектора $\boldsymbol{Y}$ на линейное подпространство, порожденное столбцами матрицы $\mathbb{X}$. Эта оценка является единственным решением уравнений, означающих ортогональность вектора остатков этому подпространству, т.е. столбцам матрицы $\mathbb{X}$:

$$\mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}}) = 0 \quad \text{или} \quad (\mathbb{X}^T\mathbb{X})\widehat{\boldsymbol{\theta}} = \mathbb{X}^T\boldsymbol{Y}, \tag{4}$$

откуда находим *МНК-оценку*

$$\widehat{\boldsymbol{\theta}} = \mathbf{B}^{-1}\mathbb{X}^T\boldsymbol{Y}. \tag{5}$$

Статистические свойства МНК-оценок

При выполнении условий $(A.1)$ и $(A.2)$ МНК-оценки обладают следующими свойствами.

1) Оценка $\widehat{\boldsymbol{\theta}}$, задаваемая формулой (5), является несмещенной, т.е. $\mathbf{E}\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

2) Матрицей ковариаций $\boldsymbol{Cov}(\widehat{\boldsymbol{\theta}}) = \|\boldsymbol{cov}(\widehat{\theta}_k, \widehat{\theta}_l)\|_{m\times m}$ служит матрица $\sigma^2\mathbf{B}^{-1}$.

3) Для любого вектора $\boldsymbol{c} \in \mathbb{R}^m$ несмещенной оценкой для величины $\boldsymbol{c}^T\boldsymbol{\theta}$ служит $\boldsymbol{c}^T\widehat{\boldsymbol{\theta}}$, причем $\mathbf{D}(\boldsymbol{c}^T\widehat{\boldsymbol{\theta}}) = \sigma^2\boldsymbol{c}^T\mathbf{B}^{-1}\boldsymbol{c}$.

4) Для любого вектора $\boldsymbol{c} \in \mathbb{R}^m$ оценка $\boldsymbol{c}^T\widehat{\boldsymbol{\theta}}$ имеет минимальную дисперсию в классе линейных (вида $\boldsymbol{d}^T\boldsymbol{Y}$) *несмещенных оценок* для $\boldsymbol{c}^T\boldsymbol{\theta}$.

Оценка остаточной дисперсии σ^2

$$\widehat{\sigma}^2 = RSS/(n-m) = F(\widehat{\boldsymbol{\theta}})/(n-m) = |\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}}|^2/(n-m)$$
$$= (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}})^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}})/(n-m) = \widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\delta}}/(n-m)$$

— несмещенная оценка параметра σ^2.

Коэффициент детерминации:

$$d = \frac{\mathbf{S}^2_{\text{о.в.}}}{\mathbf{S}^2_{\text{п.в.}}} = \frac{\mathbf{S}^2_{\text{п.в.}} - \mathbf{S}^2_{\text{ост.в.}}}{\mathbf{S}^2_{\text{п.в.}}} = 1 - \frac{\mathbf{S}^2_{\text{ост.в.}}}{\mathbf{S}^2_{\text{п.в.}}}$$

$\mathbf{S}^2_{\text{п.в.}} = \sum_{i=1}^n (Y_i - \overline{Y})^2$ — полная вариация

$\mathbf{S}^2_{\text{о.в.}} = \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$ (где $\widehat{Y}_i = \widehat{\theta}_1 X_{i,1} + \cdots + \widehat{\theta}_m X_{i,m}$) — объясненная вариация

$\mathbf{S}^2_{\text{ост.в.}} = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \widehat{\theta}_1 X_{i,1} - \cdots - \widehat{\theta}_m X_{i,m})^2 = RSS$ — остаточная вариация

Чем ближе значение d к 1, тем лучше регрессионная модель объясняет имеющиеся данные. При оценке регрессионных моделей значение d интерпретируется как соответствие модели данным. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 0.5. Модели с коэффициентом детерминации выше 0.8 можно признать достаточно хорошими. Значение

коэффициента детерминации, равное 1, означает функциональную зависимость между переменными.

НОРМАЛЬНАЯ РЕГРЕССИЯ

Для получения более содержательных заключений о распределении и свойствах оценки $\widehat{\boldsymbol{\theta}}$ предположим, что выполняется следующее предположение.

$(A.3)$ $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_{n\times n})$,

т.е. $\boldsymbol{\varepsilon}$ – сферически симметричный нормальный случайный вектор (что означает, что компоненты $\boldsymbol{\varepsilon}$ независимы, $\mathbf{E}\varepsilon_i = 0$, $\mathbf{E}\varepsilon_i^2 = \sigma^2$).

Замечание Если кроме условий $(A.1)$–$(A.2)$ выполнено еще и условие $(A.3)$, то для любого вектора $\boldsymbol{c} \in \mathbb{R}^m$ оценка $\boldsymbol{c}^T\widehat{\boldsymbol{\theta}}$ имеет минимальную дисперсию в классе ***всех несмещенных оценок*** величины $\boldsymbol{c}^T\boldsymbol{\theta}$ (а не только линейных, как в свойстве 4).

ТЕОРЕМА 1. (***Основная теорема нормальной регрессии***) В случае выполнения условий $(A.1)$–$(A.3)$ для линейной регрессионной модели (1) верны следующие утверждения.

1. Случайная величина $\sigma^{-2}RSS = \sigma^{-2}|\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\theta}}|^2$ имеет распределение хи-квадрат с $n-m$ степенями свободы и не зависит от оценки $\widehat{\boldsymbol{\theta}}$ и, поскольку $\mathbf{E}(\sigma^{-2}RSS) = n-m$, статистика $\widehat{\sigma}^2 = RSS/(n-m)$ несмещенно оценивает остаточную дисперсию σ^2.

2. Для любого вектора $\boldsymbol{c} \in \mathbb{R}^m$ случайная величина

$$\frac{\boldsymbol{c}^T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T\mathbf{B}^{-1}\boldsymbol{c}}}$$

имеет распределение Стьюдента t_{n-m} с $(n-m)$ степенями свободы.

На основе этой теоремы можно построить доверительные интервалы:
1) для σ^2,
2) для каждой из компонент θ_i параметра $\boldsymbol{\theta}$,
3) для значения $Y^* = \theta_1 X_1^* + \theta_2 X_2^* + \cdots + \theta_m X_m^*$ целевой переменной при заданном векторе значений факторов $(X_1^*, X_2^* \ldots, X_m^*)$.

Кроме того, используя критерий Стьюдента, можно проверить гипотезы о значимости каждого из факторов модели:
$H_0: \quad \theta_i = 0$
против альтернативы
$H_1: \quad \theta_i \neq 0$ для каждого $\quad i = 1, 2, \ldots, m.$

Обозначим через $\chi^2_{p_{(k)}}$ квантиль уровня p распределения хи-квадрат с k степенями свободы и через $t_{p_{(k)}}$ квантиль уровня p распределения Стьюдента k степенями свободы.

Доверительный интервал для σ^2

Пусть $1-\alpha$ – доверительная вероятность. Тогда в силу п.1 теоремы 1 имеем

$$\mathbf{P}\left(\chi^2_{\alpha/2_{(n-m)}} < \frac{RSS}{\sigma^2} < \chi^2_{1-\alpha/2_{(n-m)}}\right) = 1-\alpha$$

Отсюда следует, что с вероятностью $1-\alpha$ имеют место неравенства

$$\frac{RSS}{\chi^2_{1-\alpha/2_{(n-m)}}} < \sigma^2 < \frac{RSS}{\chi^2_{\alpha/2_{(n-m)}}}$$

Доверительный интервал для θ_i

Пусть $1-\alpha$ – доверительная вероятность. Возьмем вектор $\boldsymbol{c} = (0,0,\ldots,1,\ldots,0)^T$ с единичной i-й компонентой, остальные компоненты равны нулю. Тогда в силу п.1-2 теоремы 1 для i-й компоненты θ_i $(i = 1,2,\ldots,m)$ с вероятностью $1-\alpha$ выполняются неравенства

$$\widehat{\theta_i} - t_{1-\alpha/2_{(n-m)}}\widehat{\sigma}\sqrt{(\mathbf{B}^{-1})_{i,i}} < \theta_i < \widehat{\theta_i} + t_{1-\alpha/2_{(n-m)}}\widehat{\sigma}\sqrt{(\mathbf{B}^{-1})_{i,i}},$$

где $(\mathbf{B}^{-1})_{i,i}$ — это i-й диагональный элемент матрицы $\mathbf{B}^{-1}$.

Проверка значимости фактора X_i

Требуется проверить гипотезу
$H_0: \quad \theta_i = 0$
против альтернативы
$H_1: \quad \theta_i \neq 0.$

Если гипотеза H_0 верна, т.е. $\theta_i = 0$, то по теореме 1 величина $\frac{\theta_i}{\widehat{\sigma}\sqrt{(\mathbf{B}^{-1})_{i,i}}}$ имеет распределение Стьюдента с (n-m) степенями свободы. Отсюда правило: если

$$\left|\frac{\theta_i}{\widehat{\sigma}\sqrt{(\mathbf{B}^{-1})_{i,i}}}\right| \leq t_{1-\alpha/2_{(n-m)}},$$

что эквивалентно тому, что 0 содержится в доверительном интервале для θ_i с доверительной вероятностью $1-\alpha$, то гипотеза H_0 не отвергается на уровне значимости α, в противном случае гипотеза противоречит опытным данным. Если фактор X_i незначим (H_0 верна), то модель можно упростить, исключив фактор X_i из модели.

Доверительный интервал для прогноза $Y^* = \theta_1 X_1^* + \theta_2 X_2^* + \cdots + \theta_m X_m^*$

Теперь в качестве вектора $\boldsymbol{c}$ выберем вектор $\boldsymbol{c} = (X_1^*, X_2^* \ldots, X_m^*)^T$. Тогда для величины $\boldsymbol{c}^T\boldsymbol{\theta} = \theta_1 X_1^* + \theta_2 X_2^* + \cdots + \theta_m X_m^*$ по теореме 1 получаем доверительный интервал: с вероятностью $1-\alpha$ имеют место неравенства

$$\boldsymbol{c}^T\widehat{\boldsymbol{\theta}} - t_{1-\alpha/2_{(n-m)}}\widehat{\sigma}\sqrt{\boldsymbol{c}^T\mathbf{B}^{-1}\boldsymbol{c}} < \boldsymbol{c}^T\boldsymbol{\theta} < \boldsymbol{c}^T\widehat{\boldsymbol{\theta}} + t_{1-\alpha/2_{(n-m)}}\widehat{\sigma}\sqrt{\boldsymbol{c}^T\mathbf{B}^{-1}\boldsymbol{c}},$$

где $\boldsymbol{c}^T\widehat{\boldsymbol{\theta}} = \widehat{\theta}_1 X_1^* + \widehat{\theta}_2 X_2^* + \cdots + \widehat{\theta}_m X_m^*$ — оценка значения детерминированной составляющей Y^* в точке $X^* = (X_1^*, X_2^* \ldots, X_m^*)$.

Лабораторная работа

Рассматривается модель с четырьмя факторами:

Целевая переменная:

Y_i – рост испытуемого.

Факторы:
$X_0 = 1$ – центрирующая постоянная
X_1 – пол испытуемого
X_2 – рост отца
X_3 – рост матери
X_4 – вес испытуемого

Модель линейной регрессии

$$Y_i = \theta_0 + X_{i,1}\theta_1 + X_{i,2}\theta_2 + X_{i,3}\theta_3 + X_{i,4}\theta_4 + \varepsilon_i, \quad i = 1 \dots n, \tag{1}$$

Задание

1. Найти оценки по методу наименьших квадратов коэффициентов линейной регрессии в модели роста.

2. Найти оценку дисперсии случайной составляющей.

3. Спрогнозировать свой рост по построенной модели (по данным своих родителей).

4. Считая модель регрессии нормальной, найти доверительные интервалы (с доверительной вероятностью 0.95):
   1) для значений параметров модели (коэффициентов регрессии),
   2) дисперсии случайной составляющей,
   3) для значения прогноза (то есть для значения своего роста).

5. Проверить гипотезы о значимости каждого из факторов модели (роста отца, матери, пола, веса).

6. Вычислить коэффициент детерминации d. Насколько хорошо линейная регрессионная модель объясняет данные наблюдений?

Таблица данных

№	Y - рост студента	X_1 - пол студ.	X_2-рост отца	X_3 - рост матери	X_4 - вес студ.
1	180	1	170	170	72
2	158	0	170	155	65
3	190	1	175	195	80
4	180	1	180	170	70
5	170	1	170	165	60
6	175	0	178	179	53
7	174	1	175	163	65
8	167	0	174	167	56
9	199	1	193	173	85
10	190	1	170	166	58
11	165	1	178	156	55
12	170	1	165	160	65
13	174	0	198	176	50
14	182	1	183	167	69
15	183	1	178	172	94
16	167	0	176	164	68
17	172	0	181	187	65
18	175	1	176	164	85