

Лекция 3

Линейная регрессия, Оценки регрессионных параметров,

Лектор – *Сенько Олег Валентинович*

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

- 1 Линейная модель
- 2 Метод наименьших квадратов и его связь с методом максимального правдоподобия
- 3 Одномерная линейная модель
- 4 Многомерная линейная модель
- 5 Свойства оптимальных регрессий.
- 6 Трёхкомпонентное разложение обобщённой ошибки.
- 6 Методы, основанные на регуляризации.

Распространённым средством решения задач прогнозирования непрерывной величины Y по переменным X_1, \dots, X_n является использование метода множественной линейной регрессии. В данном методе связь переменной Y с переменными X_1, \dots, X_n задаётся с помощью линейной модели

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon, \quad (1)$$

где $\beta_0, \beta_1, \dots, \beta_n$ - вещественные регрессионные коэффициенты, ε - случайная величина, являющаяся ошибкой прогнозирования. Регрессионные коэффициенты ищутся по обучающей выборке

$$\tilde{S}_t = \{s_1 = (y_1, \mathbf{x}_1), \dots, s_m = (y_m, \mathbf{x}_m)\}, \quad (2)$$

где $\mathbf{x}_j = (x_{j1}, \dots, x_{jn})$ вектор значений переменных X_1, \dots, X_n для объекта s_j .

Традиционным способом поиска регрессионных коэффициентов является метод наименьших квадратов (МНК). МНК заключается в минимизации функционала эмпирического риска с квадратичными потерями

$$Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n) = \sum_{j=1}^m [y_j - \beta_0 - \sum_{i=1}^n x_i \beta_{ij}]^2 \quad (3)$$

То есть оценки $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n$ регрессионных коэффициентов $\beta_0, \beta_1, \dots, \beta_n$ по методу МНК удовлетворяют условию минимума функционала эмпирического риска с квадратичными потерями

$$(\hat{\beta}_0, \dots, \hat{\beta}_n) = \arg \min [Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n)]. \quad (4)$$

Предположим взаимосвязь между величиной Y и переменными X_1, \dots, X_n описывается выражением

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon_N(0, \sigma) \quad (5)$$

Где ошибка $\varepsilon_N(0, \sigma)$ распределена нормально, При 'этом дисперсия ошибки σ^2 не зависит от X_1, \dots, X_n , а математическое ожидание ошибки равно 0 при произвольных значениях прогностических переменных: $E_{\Omega}(\varepsilon_N | \mathbf{x}) = 0$, $E_{\Omega}(\varepsilon_N^2 | \mathbf{x}) = \sigma^2$ при произвольном допустимом векторе \mathbf{x} . В этом случае метод МНК тождественен более общему статистическому методу оценивания параметров статистических распределений – Методу максимального правдоподобия (ММП).

Метод максимального правдоподобия. Предположим, что некоторое пространство событий с заданной на нём вероятностной мерой P характеризуется переменными Z_1, \dots, Z_d . Метод ММП позволяет восстанавливать плотность распределения вероятностей по случайным выборкам, если общий вид плотности вероятностного распределения известен. Пусть плотность распределения принадлежит семейству функций, задаваемому вектором параметров $\theta_1, \dots, \theta_r$, принимающим значения из множества $\tilde{\Theta}$:

$$\{p(Z_1, \dots, Z_d, \theta_1, \dots, \theta_r) \mid \theta \in \tilde{\Theta}\}.$$

Предположим, что у нас имеется случайная выборка объектов, описываемых векторами z_1, \dots, z_m переменных Z_1, \dots, Z_d .

Напомним, что метод МП заключается в выборе в семействе

$$\{p(Z_1, \dots, Z_d, \theta_1, \dots, \theta_r) \mid \theta \in \tilde{\Theta}\}$$

плотности, для которой достигает максимума функция правдоподобия

$$L(\mathbf{z}_1, \dots, \mathbf{z}_m, \theta_1, \dots, \theta_r) = \prod_{j=1}^m p(\mathbf{z}_j, \theta). \quad (6)$$

Иными словами оценка $\hat{\theta}$ вектора параметров $\theta = (\theta_1, \dots, \theta_r)$ вычисляется как

$$\hat{\theta} = \arg \max_{\theta \in \tilde{\Theta}} [L(\mathbf{z}_1, \dots, \mathbf{z}_m, \theta_1, \dots, \theta_r)]. \quad (7)$$

Попытаемся вычислить значения параметров $(\beta_0, \beta_1, \dots, \beta_n)$ исходя из предположения (5). Согласно (5) разность $Y - \beta_0 - \beta_1 X_1 - \dots - \beta_n X_n$ подчиняется нормальному распределению с нулевым математическим ожиданием и дисперсией σ^2 .

Плотность распределения в пространстве переменных Y, X_1, \dots, X_n может быть восстановлена по обучающей выборке

$\tilde{S}_t = \{(y_1, \mathbf{x}_1), \dots, (y_m, \mathbf{x}_m)\}$, путём максимизации функции правдоподобия

$$L(\tilde{S}_t, \beta_0, \dots, \beta_n) = \prod_{j=1}^m \frac{1}{(\sqrt{2\pi}\sigma)} \exp \frac{-(y_j - \beta_0 - \sum_{i=1}^n \beta_i x_{ji})^2}{2\sigma^2} \quad (8)$$

Очевидно, точка экстремума функции правдоподобия $L(\tilde{S}_t, \beta_0, \dots, \beta_n)$ совпадает с точкой экстремума функции

$$\ln[L(\tilde{S}_t, \beta_0, \dots, \beta_n)] = - \sum_{j=1}^m \left[\frac{1}{2} \ln 2\pi + \ln \sigma + \frac{1}{2\sigma^2} (y_j - \beta_0 - \sum_{i=1}^n \beta_i x_{ji})^2 \right] \quad (9)$$

Однако точка максимума $\ln[L(\tilde{S}_t, \beta_0, \dots, \beta_n)]$ совпадает с точкой минимума функции $Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n)$, оптимизируемой в методе МНК, что позволяет сделать вывод о эквивалентности ММП и МНК

Рассмотрим простейший вариант линейной регрессии, описывающей связь между переменной Y и единственной переменной X :

$Y = \beta_0 + \beta_1 X + \varepsilon$. Функционал эмпирического риска на выборке $\tilde{S}_t = \{(y_1, x_1), \dots, (y_m, x_m)\}$ принимает вид

$$Q(\tilde{S}_t, \beta_0, \beta_1) = \sum_{j=1}^m [y_j - \beta_0 - x_j \beta_1]^2. \quad (10)$$

Необходимым условием минимума функционала $Q(\tilde{S}_t, \beta_0, \beta_1)$ является выполнение системы из двух уравнений

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \beta_1)}{\partial \beta_0} = -\frac{2\beta_0}{m} \sum_{j=1}^m y_j + \frac{2\beta_1}{m} \sum_{j=1}^m x_j = 0 \quad (11)$$

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \beta_1)}{\partial \beta_1} = -\frac{2\beta_0}{m} \sum_{j=1}^m x_j y_j + \frac{2\beta_1}{m} \sum_{j=1}^m x_j^2 = 0$$

Оценки $\hat{\beta}_0, \hat{\beta}_1$ являются решением системы (11) относительно параметров соответственно β_0, β_1 . Оценки регрессионных коэффициентов могут быть записаны в виде

$$\hat{\beta}_1 = \frac{\sum_{j=1}^m x_j y_j - \frac{1}{m} \sum_{j=1}^m y_j \sum_{j=1}^m x_j}{\sum_{j=1}^m x_j^2 - \frac{1}{m} (\sum_{j=1}^m x_j)^2}, \quad (12)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

, где $\bar{y} = \frac{1}{m} \sum_{j=1}^m y_j$, $\bar{x} = \frac{1}{m} \sum_{j=1}^m x_j$. Выражение для $\hat{\beta}_1$ может быть переписано в виде

$$\hat{\beta}_1 = \frac{Cov(Y, X | \tilde{S}_t)}{D(X | \tilde{S}_t)}, \quad (13)$$

где $Cov(Y, X | \tilde{S}_t)$ является выборочной ковариацией переменных Y и X , $D(X | \tilde{S}_t)$ является выборочной дисперсией переменной X .

То есть

$$\text{Cov}(Y, X | \tilde{S}_t) = \frac{1}{m} \sum_{j=1}^m (y_j - \bar{y})(x_j - \bar{x})$$

$$D(X | \tilde{S}_t) = \frac{1}{m} \sum_{j=1}^m (x_j - \bar{x})^2$$

При вычислении оценки вектора параметров $\beta = (\beta_0, \dots, \beta_n)$ в случае многомерной линейной регрессии удобно использовать матрицу плана \mathbf{X} размера $m \times (n + 1)$, которая строится по обучающей выборке \tilde{S}_t .

Матрица плана имеет вид

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ \dots & \dots & \dots & \dots \\ 1 & x_{j1} & \dots & x_{jn} \\ \dots & \dots & \dots & \dots \\ 1 & x_{m1} & \dots & x_{mn} \end{pmatrix}.$$

То есть j -я строка матрицы плана представляет собой вектор значений переменных X_1, \dots, X_n для объекта s_j с одной добавленной слева компонентой, содержащей 1.

Пусть $\mathbf{y} = (y_1, \dots, y_m)$ - вектор значений переменной Y . Связь Y с переменными X_1, \dots, X_n на объектах обучающей выборки может быть описана с помощью матричного уравнения

$$\mathbf{y} = \beta \mathbf{X}^t + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_m)$ - вектор ошибок прогнозирования для объектов \tilde{S}_t . Функционал $Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n)$ может быть записан в виде

$$Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n) = \sum_{j=1}^m [y_j - \beta_0 - \sum_{i=1}^n \beta_i \check{x}_{ji}]^2, \quad (14)$$

где \check{x}_{ji} - элементы матрицы плана \mathbf{X} , определяемые равенствами $\check{x}_{j1} = 1$, $\check{x}_{ji} = x_{j(i-1)}$ при $i > 1$.

Необходимым условием минимума функционала $Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n)$ является выполнение системы из $n + 1$ уравнений

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \dots, \beta_n)}{\partial \beta_0} = -\frac{2}{m} \left[\sum_{j=1}^m y_j \check{x}_{1j} - \sum_{j=1}^m \sum_{i=1}^{n+1} \check{x}_{ji} \check{x}_{j1} \right] = 0 \quad (15)$$

.....

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \dots, \beta_n)}{\partial \beta_n} = -\frac{2}{m} \left[\sum_{j=1}^m y_j \check{x}_{mj} - \sum_{j=1}^m \sum_{i=1}^{n+1} \check{x}_{ji} \check{x}_{jm} \right] = 0$$

Вектор оценок значений регрессионных коэффициентов $\hat{\beta}_0, \dots, \hat{\beta}_n$ является решением системы уравнений (15) .

В матричной форме система (15) может быть записана в виде

$$-2\mathbf{X}^t\mathbf{y}^t + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}^t = 0 \quad (16)$$

Решение системы (16) существует, если $\det(\mathbf{X}^t\mathbf{X}) \neq 0$. В этом случае для $\mathbf{X}^t\mathbf{X}$ существует обратная матрица и решение (16) относительно вектора $\boldsymbol{\beta}$ может быть записано в виде:

$$\hat{\boldsymbol{\beta}}^t = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}^t. \quad (17)$$

Из теории матриц следует, что $\det(\mathbf{X}^t\mathbf{X}) = 0$ если ранг матрицы \mathbf{X} по строкам менее $(n + 1)$, что происходит, если m -мерный вектор значений одной из переменных $X_{i'} \in \{X_1, \dots, X_n\}$ на выборке \tilde{S}_t является линейной комбинацией m -мерных векторов значений на \tilde{S}_t других переменных из $\{X_1, \dots, X_n\}$.

При сильной коррелированности одной из переменных $X_{i'} \in \{X_1, \dots, X_n\}$ на выборке с какой-либо линейной комбинацией других переменных значение $\det(\mathbf{X}^t\mathbf{X})$ оказывается близким к 0. При этом вычисленный вектор оценок $\hat{\beta}^t$ может сильно изменяться при относительно небольших чисто случайных изменениях вектора $\mathbf{y} = (y_1, \dots, y_m)$. Данное явление называется **мультиколлинеарностью**. Оценивание регрессионных коэффициентов с использованием МНК при наличии мультиколлинеарности оказывается неустойчивым. Отметим также, что $\det(\mathbf{X}^t\mathbf{X}) = 0$ при $n + 1 > m$. Поэтому МНК не может использоваться для оценивания регрессионных коэффициентов, когда число переменных превышает число объектов в обучающей выборке. На практике высокая устойчивость достигается только, когда число объектов в выборках по крайней мере в 3-5 раз превышает число переменных.

Рассмотрим свойства линейных регрессий, минимизирующих квадрат ошибки на пространстве событий Ω . Пусть $R(X_1, \dots, X_n)$ -регрессионная функция, которая не может быть улучшена с помощью дополнительного линейного преобразования. Иными словами при произвольных α_0 и α_1

$$E_{\Omega}[Y - \alpha_0 - \alpha_1 R]^2 \geq E_{\Omega}[Y - R]^2 \quad (18)$$

То есть минимум $E_{\Omega}[Y - \alpha_0 - \alpha_1 R]^2$ достигается при $\alpha_0 = 0$ и $\alpha_1 = 1$. Необходимым условием минимума $E_{\Omega}[Y - \alpha_0 - \alpha_1 R]^2$ является равенство 0 частных производных

$$\frac{\partial E_{\Omega}[Y - \alpha_0 - \alpha_1 R]^2}{\partial \alpha_0} = 0, \quad (19)$$

$$\frac{\partial E_{\Omega}[Y - \alpha_0 - \alpha_1 R]^2}{\partial \alpha_1} = 0.$$

Проведём простейшие преобразования

$$E_{\Omega}[Y - \alpha_0 - \alpha_1 R]^2 = E_{\Omega}Y^2 - 2\alpha_0 E_{\Omega}Y - 2\alpha_1 E_{\Omega}YR + + \\ + \alpha_1^2 E_{\Omega}R^2 + 2\alpha_0 \alpha_1 E_{\Omega}R + \alpha_0^2.$$

Откуда следует, что уравнения (19) эквивалентны уравнениям

$$2\alpha_1 E_{\Omega}R + 2\alpha_0 - 2\alpha_1 E_{\Omega}Y = 0, \quad (20)$$

$$-2E_{\Omega}(YR) + 2\alpha_1 E_{\Omega}R^2 + 2\alpha_0 E_{\Omega}R = 0.$$

Принимая во внимание, что в точке экстремума $\alpha_0 = 0$ и $\alpha_1 = 1$ получаем из системы (20) следующие свойства оптимального линейного прогнозирующего алгоритма 1) $E_{\Omega}R = E_{\Omega}Y$, 2) $E_{\Omega}(YR) = E_{\Omega}R^2$,

Из свойств 1) 2) следует , что дисперсия R равна ковариации Y и R .
Действительно,

$$D(R) = E_{\Omega}(R - E_{\Omega}R)^2 = E_{\Omega}R^2 - (E_{\Omega}R)^2$$

. Однако

$$cov(Y, R) = E_{\Omega}(R - E_{\Omega}R)(Y - E_{\Omega}Y) = E_{\Omega}YR - E_{\Omega}R^2$$

. То есть

$$3)cov(Y, R) = D(R).$$

Рассмотрим теперь коэффициент корреляции между Y и R - $\rho(Y, R)$.

$$\rho(Y, R) = \frac{cov(Y, R)}{\sqrt{D(Y)D(R)}} = \sqrt{\frac{D(R)}{D(Y)}}.$$

Рассмотрим величину ошибки прогнозирования Y с помощью R .

$$\begin{aligned} 4) \Delta(Y, R) &= E_{\Omega}(Y - R)^2 = E_{\Omega}Y^2 - 2E_{\Omega}(YR) + E_{\Omega}R^2 = \\ &= E_{\Omega}Y^2 - E_{\Omega}R^2 = E_{\Omega}Y^2 - (E_{\Omega}Y)^2 + (E_{\Omega}Y)^2 - E_{\Omega}R^2 = \\ &= E_{\Omega}Y^2 - E_{\Omega}R^2 = E_{\Omega}Y^2 - (E_{\Omega}Y)^2 + (E_{\Omega}R)^2 - E_{\Omega}R^2 = D(Y) - D(R). \end{aligned}$$

Из свойств (3) и (4) легко следует свойство для относительной ошибки $\Delta_{rel}(Y, R) = \frac{\Delta(Y, R)}{D(Y)}$:

$$5) \Delta_{rel}(Y, R) = 1 - \rho^2(Y, R).$$

Напомним, что обобщающая способность алгоритма прогнозирования A , обученного по выборке \tilde{S}_t с помощью некоторого метода M измеряется величиной потерь на генеральной совокупности Ω :

$$E_{\Omega} \lambda[Y, A(\mathbf{x}, \tilde{S}_t)].$$

Для оценки эффективности использования метода прогнозирования M для прогнозирования случайного процесса, связанного с генеральной совокупностью Ω , при фиксированном размере обучающей выборки m естественно использовать математическое ожидание потерь по пространству всевозможных обучающих выборок длины m :

$$\Delta_G = E_{\Omega_m} E_{\Omega} \lambda[Y, A(\mathbf{x}, \tilde{S}_t)],$$

где Ω_m - рассмотренное ранее пространство обучающих выборок длины m .

Величину Δ_G будем называть **обобщёнными потерями**. При использовании в качестве функции потерь квадрата ошибки обобщённые потери становятся **обобщённой квадратичной ошибкой** и принимают вид

$$\Delta_G = E_{\Omega_m} E_{\Omega} [Y - A(\mathbf{x}, \tilde{S}_t)]^2.$$

Проведём простые преобразования:

$$\begin{aligned} E_{\Omega_m} E_{\Omega} [Y - A(\mathbf{x}, \tilde{S}_t)]^2 &= E_{\Omega_m} E_{\Omega} [Y - E(Y | \mathbf{x}) + \\ &+ E(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2 = E_{\Omega_m} E_{\Omega} [Y - E(Y | \mathbf{x})]^2 + \\ &+ E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2 + \\ &+ 2E_{\Omega_m} E_{\Omega} \{ [E(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] [Y - E(Y | \mathbf{x})] \}. \end{aligned}$$

Покажем справедливость равенства

$$E_{\Omega_m} E_{\Omega} \{ [E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] [Y - E(Y | \mathbf{x})] \} = 0 \quad (21)$$

Действительно

$$E_{\Omega} \{ [E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] [Y - E(Y | \mathbf{x})] \} =$$

$$= \int_{\tilde{X}} E_{\Omega} \{ [E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] [Y - E(Y | \mathbf{x})] | \mathbf{x} \} p(\mathbf{x}) dX_1 \dots dX_n,$$

где \tilde{X} совместная область допустимых значений X_1, \dots, X_n в пространстве \mathbb{R}^n , $p(\mathbf{x})$ - плотность вероятности внутри \tilde{X} . При любом фиксированном \mathbf{x} множитель $[E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]$ не зависит от объекта, для которого производится прогнозирование, и может быть вынесен за знак математического ожидания:

$$\begin{aligned} E_{\Omega}\{[E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)][Y - E(Y | \mathbf{x})] | \mathbf{x}\} &= \\ &= [E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]E_{\Omega}\{[Y - E_{\Omega}(Y | \mathbf{x})] | \mathbf{x}\}. \end{aligned}$$

Однако по определению условного математического ожидания при любом \mathbf{x}

$$E_{\Omega}\{[Y - E_{\Omega}(Y | \mathbf{x})] | \mathbf{x}\} = 0.$$

Откуда следует справедливость равенства (21). Принимая во внимание, что $[Y - E_{\Omega}(Y | \mathbf{x})]$ не зависит от \tilde{S}_t получаем

$$E_{\Omega_m}E_{\Omega}[Y - E(Y | \mathbf{x})]^2 = E_{\Omega}[Y - E(Y | \mathbf{x})]^2$$

. В итоге

$$\Delta_G = E_{\Omega}[Y - E(Y | \mathbf{x})]^2 + E_{\Omega_m}E_{\Omega}[E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2.$$

Введём обозначение

$$\hat{A}(\mathbf{x}) = E_{\Omega_m}A(\mathbf{x}, \tilde{S}_t)$$

Компонента разложения

$$E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2$$

может быть представлена в виде

$$\begin{aligned} & E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x}) + \hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2 = \\ & = E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})]^2 + E_{\Omega_m} E_{\Omega} [\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2 + \\ & \quad + 2E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})][\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] \end{aligned}$$

Справедливо равенство

$$2E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})][\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] = 0. \quad (22)$$

Действительно

$$\begin{aligned} & 2E_{\Omega_m} E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})][\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] = \\ & 2E_{\Omega} \{ [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})] E_{\Omega_m} [\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] \}. \end{aligned}$$

Однако из определения $\hat{A}(\mathbf{x})$ следует, что

$$E_{\Omega} [\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)] = 0.$$

Поэтому равенство (22) справедливо. В итоге справедливо трёхкомпонентное разложение обобщённой квадратичной ошибки Δ_G :

$$\begin{aligned} \Delta_G &= E_{\Omega} [Y - E(Y | \mathbf{x})]^2 + E_{\Omega} [E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})]^2 + \quad (23) \\ &+ E_{\Omega_m} E_{\Omega} [\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2 = \\ &= \Delta_N + \Delta_B + \Delta_V \end{aligned}$$

Шумовая компонента.

$$\Delta_N = E_{\Omega}[Y - E(Y | \mathbf{x})]^2$$

является минимально достижимой квадратичной ошибкой прогноза, которая не может быть устранена с использованием только математических средств. Составляющая сдвига (Bias).

$$\Delta_B = E_{\Omega}[E_{\Omega}(Y | \mathbf{x}) - \hat{A}(\mathbf{x})]^2$$

Высокое значение компоненты сдвига соответствует отсутствию в модели $\tilde{M} = \{A : \tilde{X} \rightarrow \tilde{Y}$, внутри которой осуществляется поиск, алгоритмов, достаточно хорошо аппроксимирующих объективно существующую зависимость Y от переменных X_1, \dots, X_n .

Составляющая сдвига может быть снижена, например, путём расширения модели за счёт включения в него дополнительных более сложных алгоритмов, что обычно позволяет повысить точность аппроксимации данных.

Дисперсионная составляющая (Variance).

$$\Delta_V = E_{\Omega_m} E_{\Omega} [\hat{A}(\mathbf{x}) - A(\mathbf{x}, \tilde{S}_t)]^2$$

характеризует неустойчивость обученных прогнозирующих алгоритмов при статистически возможных изменениях в обучающих выборках. Дисперсионная составляющая возрастает при небольших размерах обучающей выборки. Дисперсионная составляющая может быть снижена путём выбора сложности модели, соответствующей размеру обучающих данных.

Таким образом существует **Bias-Variance дилемма** Составляющая сдвига может быть снижена путём увеличения разнообразия модели. Однако увеличение разнообразия модели при недостаточном объёме обучающих данных ведёт к росту компоненты сдвига. Наиболее высокая точность прогноза достигается, при поддержании правильного баланса между разнообразием используемой модели и объёмом обучающих данных

Одним из возможных способов борьбы с неустойчивостью является использование методов, основанных на включение в исходный оптимизируемый функционал $Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n)$ дополнительной штрафной компоненты. Введение такой компоненты позволяет получить решение, на котором $Q(\tilde{S}_t, \beta_0, \beta_1, \dots, \beta_n)$ достаточно близок к своему глобальному минимуму. Однако данное решение оказывается значительно более устойчивым и благодаря устойчивости позволяет достигать существенно более высокой обобщающей способности. Подход к получению более эффективных решений с помощью включения штрафного слагаемого в оптимизируемый функционал принято называть **регуляризацией по Тихонову**.

На первом этапе переходим от исходных переменных $\{X_1, \dots, X_n\}$ к стандартизированным $\{X_1^s, \dots, X_n^s\}$, где $X_i^s = \frac{X_i - \hat{X}_i}{\hat{\sigma}_i}$,

$\hat{X}_i = \frac{1}{m} \sum_{j=1}^m x_{ji}$, $\sqrt{\frac{1}{m} \sum_{j=1}^m (X_i - x_{ji}^2)}$ а также от исходной прогнозируемой переменной Y к стандартизованной прогнозируемой переменной $Y^s = Y - \frac{1}{m} \sum_{j=1}^m y_j$. Пусть $\check{x}_{j1}^s = 1$, $\check{x}_{ji}^s = x_{j(i-1)}^s$ при $i > 1$, где $x_{j(i-1)}^s$ - значение признака X_i^s для j-го объекта. Пусть

также $\mathbf{X}_s = \begin{pmatrix} 1 & x_{11}^s & \dots & x_{1n}^s \\ \dots & \dots & \dots & \dots \\ 1 & x_{j1}^s & \dots & x_{jn}^s \\ \dots & \dots & \dots & \dots \\ 1 & x_{m1}^s & \dots & x_{mn}^s \end{pmatrix}$. - матрица плана для

стандартизованных переменных, $\mathbf{y}^s = (y_1^s, \dots, y_m^s)$ - вектор значений стандартизованной переменной Y_s .

Одним из первых методов регрессии, использующих принцип регуляризации, является метод **гребневой регрессии (ridge regression)**. В гребневой регрессии в оптимизируемый функционал дополнительно включается сумма квадратов регрессионных коэффициентов при переменных $\{X_1^s, \dots, X_n^s\}$. В результате функционал имеет вид

$$Q_{\text{ridge}}(\tilde{S}_t, \beta_0, \dots, \beta_n) = \sum_{j=1}^m [y_j - \beta_0 - \sum_{i=1}^n \beta_i \tilde{x}_{ji}^s]^2 + \gamma \sum_{i=1}^n \beta_i^2, \quad (24)$$

где γ - положительный вещественный параметр. Пусть β_r является вектором оценок регрессионных коэффициентов, полученным в результате минимизации $Q_{\text{ridge}}(\tilde{S}_t, \beta_0, \dots, \beta_n)$.

Отметим, что увеличение регрессионных коэффициентов приводит к увеличению $Q_{ridge}(\tilde{S}_t, \beta_0, \dots, \beta_n)$. Таким образом использование гребневой регрессии приводит к снижению длины вектора регрессионных коэффициентов при переменных $\{X_1^s, \dots, X_n^s\}$. Рассмотрим конкретный вид вектора регрессионных коэффициентов β_r в гребневой регрессии. Необходимым условием минимума функционала $Q_{ridge}(\tilde{S}_t, \beta_0, \dots, \beta_n)$ является выполнение системы из $n + 1$ уравнений:

$$\frac{\partial Q_{ridge}(\tilde{S}_t, \beta_0, \dots, \beta_n)}{\partial \beta_0} = -\frac{2}{m} \left[\sum_{j=1}^m y_j \check{x}_{1j}^s - \sum_{j=1}^m \sum_{i=1}^{n+1} \check{x}_{ji}^s \check{x}_{j1}^s + \gamma \beta_0 \right] = 0 \quad (25)$$

.....

$$\frac{\partial Q_{ridge}(\tilde{S}_t, \beta_0, \dots, \beta_n)}{\partial \beta_n} = -\frac{2}{m} \left[\sum_{j=1}^m y_j \check{x}_{mj}^s - \sum_{j=1}^m \sum_{i=1}^{n+1} \check{x}_{ji}^s \check{x}_{jm}^s + \gamma \beta_n \right] = 0$$

Поэтому вектор оценок регрессионных коэффициентов в методе гребневая регрессия является решением системы (25).

В матричной форме система (25) может быть записана в виде

$$-2\mathbf{X}_s^t \mathbf{y}_s^t + (2\mathbf{X}_s^t \mathbf{X}_s + 2\gamma \mathbf{I}) \boldsymbol{\beta}_r^t = 0 \quad (26)$$

или в виде

$$\boldsymbol{\beta}_r^t = \mathbf{X}_s^t \mathbf{y}_s^t (\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I})^{-1} \quad (27)$$

где \mathbf{I} - единичная матрица. Отметим, что произведение $\mathbf{X}_s^t \mathbf{X}_s$ представляет собой симметрическую неотрицательно определённую матрицу. Матрица $\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I}$ также является симметрической матрицей. Каждому собственному значению λ_k матрицы $\mathbf{X}_s^t \mathbf{X}_s$ соответствует собственное значение $\lambda_k + \gamma$ матрицы $\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I}$. Пусть λ_{min}^γ минимальное собственное значение матрицы $\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I}$ удовлетворяет неравенству $\lambda_{min}^\gamma > \gamma$. Откуда следует, что всегда $\det(\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I}) > 0$, а обратная матрица $(\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I})^{-1}$ всегда существует. Достаточно большая величина $\det(\mathbf{X}_s^t \mathbf{X}_s + \gamma \mathbf{I})$ приводит к относительно небольшим изменениям оценок регрессионных коэффициентов при небольших изменениях в обучающих выборках.

Наряду с гребневой регрессией в последние годы получил распространение **метод Лассо**, основанный на минимизации функционала

$$Q_{lasso}(\tilde{S}_t, \beta_0, \dots, \beta_n) = \sum_{j=1}^m [y_j - \beta_0 - \sum_{i=1}^n \beta_i \tilde{x}_{ji}^s]^2 + \gamma \sum_{i=1}^n |\beta_i|. \quad (28)$$

Интересной особенностью метода Лассо является равенство 0 части из регрессионных коэффициентов. Однако равенство 0 коэффициента на самом деле означает исключение из модели соответствующей ему переменной. Поэтому метод Лассо не только строит оптимальную регрессионную модель, но и производит отбор переменных. Метод может быть использован для отбора переменных в условиях, когда размерность данных превышает размер выборки. Отметим, что общее число отобранных переменных не может превышать размера обучающей выборки. Эксперименты показали, что эффективность отбора переменных методом Лассо снижается, при высокой взаимной корреляции некоторых из них.

Данными недостатками не обладает другой метод построения регрессионной модели, основанный на регуляризации по Тихонову, который называется **эластичная сеть**. Метод эластичная сеть основан на минимизации функционала

$$Q_{elnet}(\tilde{S}_t, \beta_0, \dots, \beta_n) = \sum_{j=1}^m [y_j - \beta_0 - \sum_{i=1}^n \beta_i \tilde{x}_{ji}^s]^2 + \gamma \sum_{i=1}^n [\beta_i^2 \theta + (1 - \theta) |\beta_i|], \quad (29)$$

где $\theta \in [0, 1]$. Метод эластичная сеть включает в себя метод гребневая регрессия и Лассо как частные случаи.