Лекция 2.2

Корреляционный и регрессионный анализ

Регрессионный анализ

Регрессионный анализ – это инструмент для количественного определения значения одной переменной на основании другой.

Парная (простая) линейная регрессия даёт нам правила, определяющие линию регрессии, которая лучше других предсказывает наиболее вероятные значения одной переменной на основании другой (переменных всего две).

Множественная регрессия является расширением простой линейной регрессии.

- ***Для простого регрессионного анализа*** предполагается:
  для любого произвольного или фиксированного значения X соответствующая ему величина Y имеет нормальное распределение относительно некоторого теоретического среднего значения.
- График зависимости этих средних от X отражает основное соотношение между X и Y.
- Зависимая и независимые переменные должны быть измерены в ***метрической шкале.***

Регрессионный анализ

- Коэффициент при независимой переменной X в уравнении регрессии (a_1) называется ***коэффициентом регрессии*** Y на X.
- Он определяет угол наклона прямой на графике и служит мерой среднего изменения величины Y при изменении X на единицу.
- Коэффициент регрессии может быть положительным и отрицательным, а если X и Y независимы, то он равен нулю.

Общая задача регрессионного анализа. МНК

- Общая задача регрессионного анализа состоит в том, чтобы
  - по наблюдениям x_i и y_i оценить параметры модели a_0 и a_1 «*наилучшим образом*»;
  - проверить гипотезу о значимости уравнения и коэффициентов регрессии;
  - оценить адекватность полученной зависимости и т.д.
- Если под «наилучшим образом» понимать минимальную сумму квадратов расстояний до прямой от наблюдаемых точек, вычисленных вдоль оси ординат, то такой метод построения уравнения регрессии называется ***методом наименьших квадратов***.

- Найдем теперь оценку неизвестных значений a_0 и a_1, основанную на имеющейся у нас выборке объема n. *Наилучшие* оценки b_0 и b_1 для a_0 и a_1 получаются *минимизацией* соответственно по a_0 и a_1 суммы квадратов отклонений

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2.$$

- Как известно, минимум функции можно найти, приравняв к нулю ее производную.
- Далее находим частные производные функции S по a_0 и a_1 и приравниваем их к нулю.

- Решая полученную систему уравнений находим оценки наименьших квадратов:

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- Оценкой уравнения регрессии (или прямой наименьших квадратов) будет

$$\hat{y} = b_0 + b_1 x.$$

- Разницей между наблюдаемым и предсказанным значением Y при X=x_i называется отклонением или остатком: $d_i = y_i - \widehat{y}_i$.

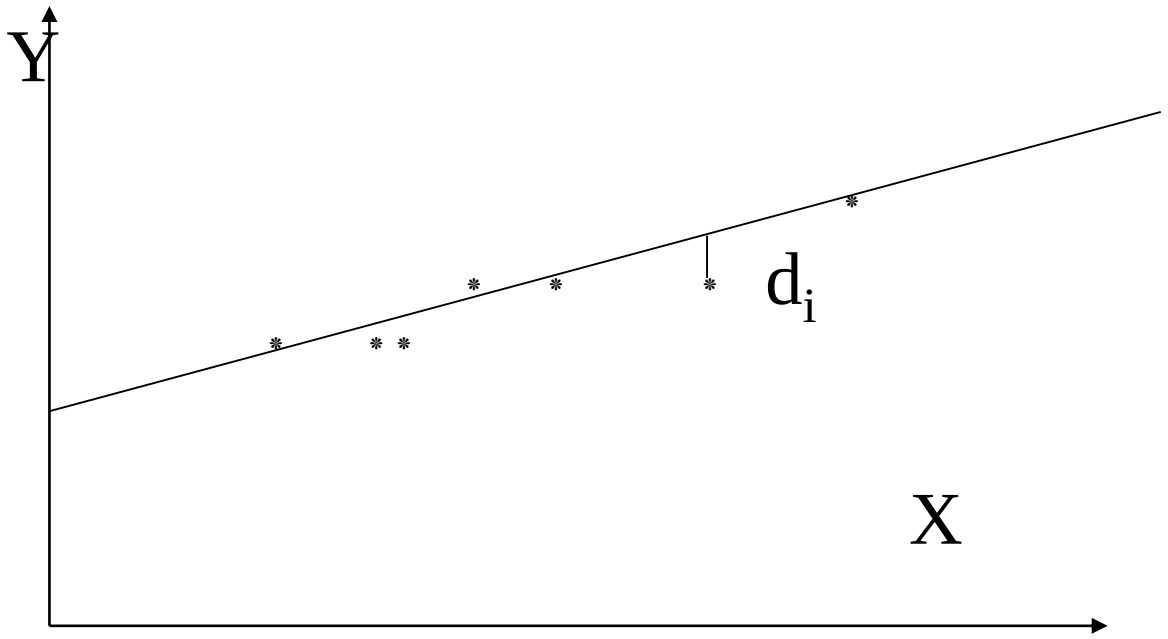

Рис. График прямой наименьших квадратов

- Рассмотрим некоторые свойства полученных оценок b_0 и b_1:
- 1) b_0, b_1 независимые случайные величины;
- 2) оценки подчиняются нормальному закону распределения;
- 3) математическое ожидание $M(b_0)=a_0$ и $M(b_1)=a_1$;
- 4) ковариация $cov(b_0,b_1)=0$;
- 5) дисперсии оценок равны

$$\sigma_{b_0}^2 = \sigma^2 / n, \qquad \sigma_{b_1}^2 = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где σ^2 – дисперсия ошибок.

- Чтобы сделать статистические выводы о b_0, b_1 и $\hat{y}$, сначала необходимо оценить дисперсию σ^2. Согласно теории общей линейной модели, обычная несмещенная оценка для σ^2 определяется через дисперсию оценки

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2}{n-2}.$$

- Положительный квадратный корень из этой величины называют ***стандартной ошибкой оценки***.

- Для проверки нулевой гипотезы о том, что простая линейная регрессия Y по X отсутствует (H_0: $b_1=0$), построим таблицу дисперсионного анализа.
- Если полученное значение F-отношения больше табличного с заданным уровнем значимости α и числом степенями свободы (1; n-2), то гипотезу отвергаем ($b_1 \neq 0$), т.е. Y линейно зависит от X.

Таблица дисперсионного анализа для модели линейной регрессии

Источник дисперсии	Сумма квадратов	Степени свободы	Средний квадрат	F-отношение
Регрессия	$SS_D = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$	$v_D=1$	$MS_D=SS_D/v_D$	$F = MS_D / MS_R$
Отклонение от регрессии	$SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$v_R=n-2$	$MS_R=s^2= SS_R/v_R$	
Полная	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$v_T=n-1$		

Примечание. SS_D – обусловленная регрессией сумма квадратов; SS_R – сумма квадратов отклонений от линии регрессии или остаточная сумма квадратов; SS_T – полная сумма квадратов.

Регрессионный анализ. Коэффициент детерминации

- Отношение SS_D/SS_T есть доля вариации Y, объясняемая регрессией Y по X. Это отношение называется ***коэффициентом детерминации*** (r^2).
- Коэффициент детерминации является мерой качества предсказанных значений зависимой переменной Y моделью линейной регрессии.
- Интервал распределения коэффициента детерминации - [0;1].
- Если $r^2=1$, то наблюдаемые точки в точности лежат на линии регрессии.
- Если $r^2=0$, то Y не зависит от X.

- Если, например, r^2=0,95 , это означает, что 95% отклонений от среднего значения зависимой переменной объясняет построенная регрессия, а 5% отклонений остаются необъясненными.

- Далее в регрессионном анализе для проверки адекватности полученной модели проводят анализ остатков. Для этого строят график d_i в зависимости от x_i или $\hat{y}_i$, i=1,…n.

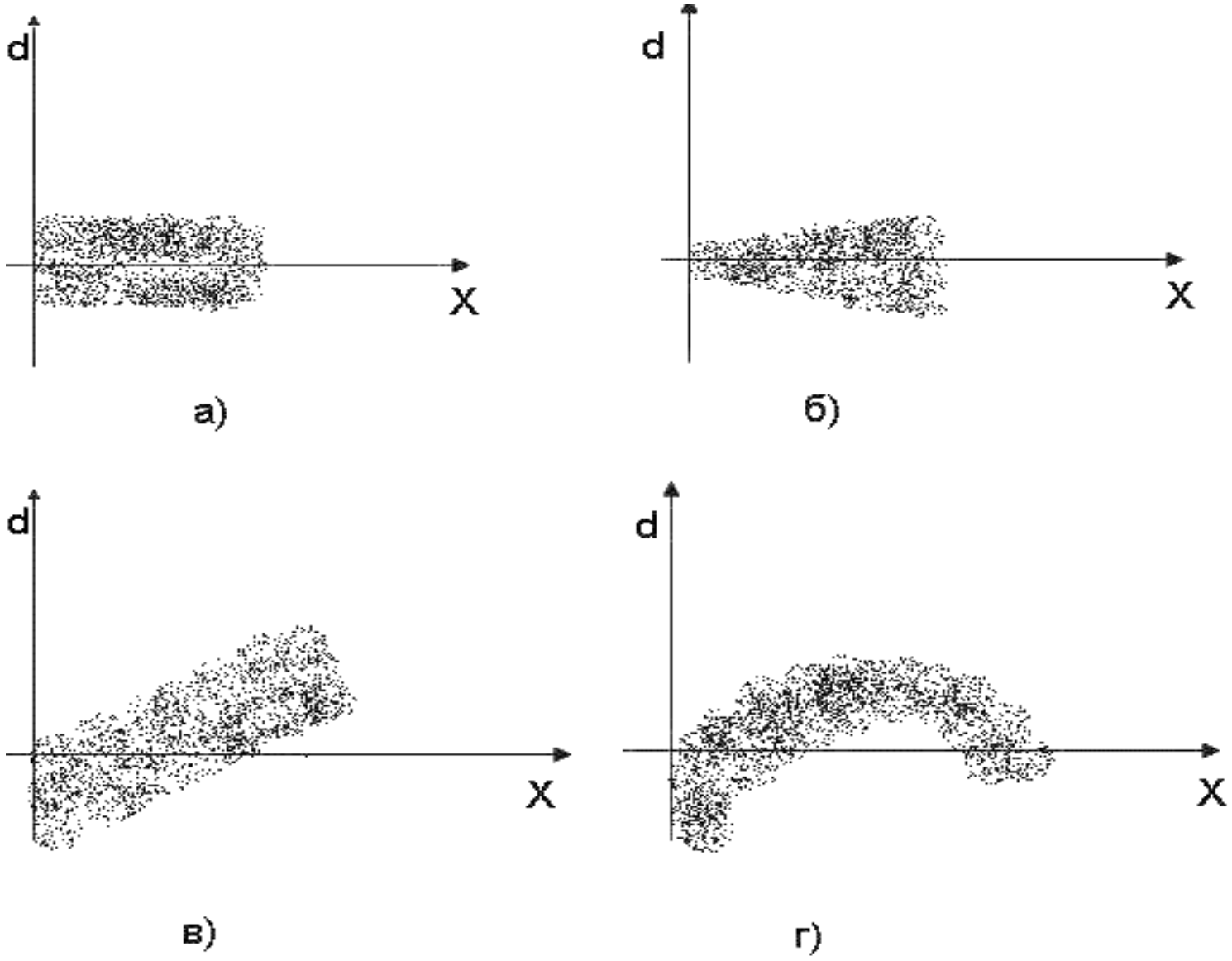

Рис. Примеры графиков остатков

- Если остатки попадают в горизонтальную полосу с центром на оси абсцисс, модель можно рассматривать как ***адекватную*** (*рис.а)*.
- Если полоса расширяется, когда x или $\hat{y}$ возрастает (рис. б), это указывает на ***гетероскедастичность*** (т.е. на отсутствие постоянства дисперсии σ^2).
- В частности, σ может быть функцией $\beta_0+\beta_1 x$, что делает необходимым преобразование переменной Y.
- График, показывающий ***линейный тренд*** (рис. в), дает основание для введения в модель дополнительной независимой переменной.
- График вида, представленного на рисунке г), указывает, что в модель должен быть добавлен линейный или квадратный член.

- Если предсказанная регрессия удовлетворительно описывает истинную зависимость между Y и X, то остатки должны быть независимыми нормально распределенными случайными величинами с нулевым средним, и в значениях d_i должен отсутствовать тренд.

- **Независимость остатков** может быть проверена при помощи ***коэффициента Дарбина-Ватсона***, имеющего вид:

$$D = \sum_{i=2}^{n} (d_i - d_{i-1})^2 \Big/ \sum_{i=1}^{n} d_i^2 .$$

- Если $D>D_1$, то с достоверностью α принимается гипотеза о наличии соответственно отрицательной или положительной корреляции остатков. Если
- $D_2(\alpha)>D>D_1(\alpha)$, то критерий не позволяет принять решение по гипотезе о наличии или отсутствии корреляции остатков.
- Если $D_2(\alpha)<D<4 - D_2(\alpha)$, то гипотеза о корреляции остатков отклоняется. Критические значения $D_1(\alpha)$ и $D_2(\alpha)$ для различных α берутся из табличных данных.

Итак, определим **основные этапы регрессионного анализа**:

1) нахождение коэффициентов регрессии, построение модели;

2) проверка гипотезы о существовании линейной зависимости между переменными;

3) анализ остатков.

Множественная регрессия

Уравнение регрессии — это зависимость случайной величины Y от неслучайных факторов ***X****, т. е. зависимость «следствия» Y от «причин»* ***X****:*

$$Y=\eta\,(\boldsymbol{X},\boldsymbol{\beta})+\boldsymbol{\varepsilon}, \qquad (2.1)$$

где $\boldsymbol{X}=\{x_1, x_2,\ldots,x_j,\ldots,x_k\}$ - вектор факторов, $j= 1,2,\ldots, k$;

$\boldsymbol{\beta}=\{\beta_1,\beta_2,\ldots,\beta_j,\ldots,\beta_d\}$ - вектор параметров модели;

$\eta(\boldsymbol{X},\boldsymbol{\beta})$ -функция регрессии (или функция отклика) случайной величины $\boldsymbol{Y}$ на неслучайные $\boldsymbol{X}$;

$\boldsymbol{\varepsilon}=\{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_i,\ldots,\varepsilon_n\}$ - вектор ошибок наблюдений.

При многократном однотипном воздействии $\boldsymbol{X}$ на входе получаем на выходе объекта различные значения Y.

Множественная регрессия

- Уравнение регрессии вида (2.1) описывает только статику объекта, т. е. предполагается, что взаимосвязь показателя ***Y*** и факторов ***X***, установленная в определенный момент (интервал) времени, от времени не зависит.

- ***Регрессионные модели в зависимости от рассматриваемых факторов могут быть использованы в целях***: объяснения сути явления (предсказательная модель), прогнозирования (прогнозная модель), управления.

- Если установлена зависимость *Y* только от управляемых факторов ***X'***, то это уравнение теоретически может быть использовано в целях управления объектом.

Множественная регрессия

- При построении уравнения по результатам пассивного эксперимента ошибка в управлении может быть неприемлемой.
- Функциональная модель и модель для прогнозирования содержат все группы факторов $\boldsymbol{X}'$, $\boldsymbol{Z}$:

$$Y=\eta(\boldsymbol{X}', \boldsymbol{Z}, \boldsymbol{\beta})+\varepsilon.$$

Обычно функциональная модель более сложная, чем предсказательная.

- В целях построения $\eta(\boldsymbol{X},\boldsymbol{\beta})$ обычно предполагают, что это - гладкая функция в области допустимых значений: $\boldsymbol{X} \in \boldsymbol{X}_{доп}$.
- В этом случае возможно ее разложение в ряд Тейлора в окрестности некоторой точки, например, точки, соответствующие «центру» эксперимента, - среднему значению $\bar{X}$.

Множественная регрессия

В результате получаем полином степени p вида:

$$Y = \beta_0 + \sum_j \beta_j x_j + \sum_{uj} \beta_{uj} x_u x_j + \sum_j \beta_{jj} x_j^2 + \cdots + \varepsilon, \quad (2.2)$$

где $\sum_j$ – сумма по j = 1,2,…,k,

$\sum_{uj}$ – сумма парных взаимодействий $x_u x_j$, u,j =1,2,…,k, $u \neq j$, k - число факторов;

β_{uj} - коэффициент парного взаимодействия,

β_{jj} - коэффициент при квадрате переменной и т. д.;

в формуле (2.2) степень полинома p = 2.

- Обычно разложение ограничивают конечным числом членов ряда. Например:

$$Y = \beta_0 + \sum_j \beta_j x_j \qquad (2.3)$$

Множественная регрессия

- По результатам эксперимента могут быть определены не «истинные» коэффициенты регрессии **β**, соответствующие генеральной совокупности, а лишь их оценки

$$\boldsymbol{B} = (b_0, b_1, \ldots, b_j, \ldots, b_d),$$

вычисленные по выборке объемом **n**.

- В этом случае уравнение регрессии в векторной форме имеет вид:

$$\hat{Y} = \eta(\boldsymbol{X}, \boldsymbol{B}), \quad (2.4)$$

где $\hat{Y}$ – предсказанные (прогнозируемые) значения выходной величины.

- При выводе и использовании формул регрессионного анализа удобнее пользоваться векторной формой представления уравнений регрессии:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}; \quad \hat{\boldsymbol{Y}} = \boldsymbol{XB}, \quad (2.5)$$

Множественная регрессия

Y – вектор наблюдений; ***X***- матрица значений независимых переменных;

β, B - векторы коэффициентов и их оценок соответственно;

ε - вектор ошибок:

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_i \\ \cdot \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\mathrm{j}} & \cdots & x_{1\mathrm{k}} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\mathrm{j}} & \cdots & x_{2\mathrm{k}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{\mathrm{i}1} & x_{\mathrm{i}2} & \cdots & x_{\mathrm{ij}} & \cdots & x_{\mathrm{ik}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n\mathrm{j}} & \cdots & x_{\mathrm{nk}} \end{bmatrix}, B=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \beta_i \\ \cdot \\ \beta_d \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \varepsilon_i \\ \cdot \\ \varepsilon_n \end{bmatrix}.$$

Первый столбец матрицы ***X*** содержит фиктивную переменную $x_{0i}=1, i=1,2,\ldots,n$.

Множественная регрессия

В общем случае, когда $d > k$ (число коэффициентов регрессии больше числа анализируемых факторов k), можно записать уравнение регрессии в следующей векторной форме:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{FB}, \quad (2.6)$$

где $F[f_{iq} \ (X)]_{nd}$ - матрица известных функций f_{iq}, от независимых переменных.

Множественная регрессия

Например, пусть $k = 2$, а $d = 5$, т. е. необходимо вычислить пять коэффициентов: b_0, b_1, b_2, b_3, b_4.

Тогда:

$F = (f_{i1}, f_{i2}, f_{i3}, f_{i4}, f_{i5})$, где $f_{i0} = 1 = x_{i0}, f_{i1} = x_{i1}, f_{i2} = x_{i2}, f_{i3} = x_{i1}x_{i2}, f_{i4} = x_{i1}, f_{i5} = x_{i2},$

т. е. введены переобозначения, и вместо уравнения

$$\hat{Y} = b_0 + b_1x_1 + b_2x_2\, b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2$$

имеем

$$\hat{Y} = b_0 + b_1f_1 + b_2\, f_2 + b_3f_3 + b_4\, f_4 + b_5\, f_5 \quad (2.7)$$

После вычисления коэффициентов регрессии нужно вернуться к первоначальным обозначениям, для того чтобы облегчить интерпретацию результатов.

Задачи регрессионного анализа:

- вычисление коэффициентов регрессии;
- проверка значимости коэффициентов регрессии;
- проверка адекватности модели;
- выбор «лучшей» регрессии;
- вычисление стандартных ошибок.

Вычисление коэффициентов регрессии осуществляется методом наименьших квадратов (МНК-метод).

Проверка значимости коэффициентов регрессии основана на методах проверки «гипотез о средних». Проверка адекватности модели основана на методах дисперсионного анализа.

Вычисление стандартных ошибок, по которым можно судить о точности предсказаний, осуществляется по обычным формулам расчета средних квадратичных отклонений.

Постулаты регрессионного анализа

- ***Первое условие.*** **Результаты эксперимента должны быть свободны от систематических ошибок, т. е. ожидание *M{Y}* величины *Y* должно быть равно действительному значению $\tilde{Y}$**

т.е.: $M\{Y\} = \tilde{Y}, \qquad \tilde{Y} = (\tilde{Y}_1, \dots, \tilde{Y}_i, \dots, \tilde{Y}_N).$

Следовательно, математическое ожидание ошибки ε будет равно нулю

$$M\{\varepsilon\} = M\{Y - \tilde{Y}\} = 0,$$

или если действительным значением считать предсказанное по уравнению регрессии значение $\hat{Y}$, то:

$$M\{\varepsilon\} = M\{Y - \hat{Y}\} = 0,$$

Рассмотрим отдельный опыт в точке x_i. Пусть для «истинной модели» (2.1) величина среднего $M\{\ddot{Y}_i\}$ значения $\tilde{y}_i$ равна:

$$M\{\tilde{Y}_i\} = \tilde{y}_i$$

Определим ошибку ε_i i-го опыта:

$$\varepsilon_i = Y_i - \hat{Y} = \left[\left(Y_i - \hat{Y}_i\right) - \left(\tilde{Y}_i - M\{\tilde{Y}_i\}\right)\right] + [\tilde{Y}_i - M\{\tilde{Y}_i\}] = A_i + B_i,$$

где $A_i = (Y_i - \hat{Y}_i) - (\ddot{Y}_i - M\{\hat{Y}_i\})$ - случайная переменная с нулевым средним;

$B_i = [\tilde{Y}_I - M\{\tilde{Y}_I\}]$ - ошибка смещения.

Если построенная модель верна (корректна), то ошибка смещения равна нулю, и первое условие соблюдено.

Постулаты регрессионного анализа

- ***Второе условие*** **- дисперсия результатов наблюдения во всех лоточках одинакова**, т. е.:

$$D\{Y_i\} = \sigma^2, D\{\varepsilon_i\} = \sigma^2 \text{ для } \forall i.$$

- ***Третье условие*** **- результаты наблюдений в точке x_i не зависят от результатов наблюдений в предыдущей точке x_{i-1}**, т.е. Y_{i-1} и Y_i - не коррелированны, так что ковариации равны нулю:

$$Cov\{Y_{i-1}, Y_i\} = M\{(Y_{i-1} - \hat{Y}_{i-1})(Y_i - \hat{Y}_i)\} = 0;$$
$$Cov\{\varepsilon_{i-1}, \varepsilon_i\} = M\{\varepsilon_{i-1}, \varepsilon_i\} = 0;$$

Поэтому для уравнения регрессии имеем, например:

$$M\{Y_i\} = \beta_0 + \sum_j \beta_j x_{ij} + \sum_{uj} \beta_{uj} x_{iu} x_{ij} + \sum_j \beta_{jj} x_{ij}^2 + \cdots + \varepsilon_i,$$

- ***Четвертое условие:*** **Y_i, ε_i - случайные величины, подчиненные нормальному закону распределения** со средними

$M\{\tilde{Y}_i\}$ *и дисперсиями* $D\{\tilde{Y}_i\} = \sigma^2$,*т. е.*

$$Y_i \approx N(\tilde{Y}_i, \sigma^2);$$
$$\varepsilon_i \approx N(0, \sigma^2),$$

где N – обозначение нормальных распределений наблюдаемой величины Y и ее ошибки ε.

Постулаты регрессионного анализа

Представленные условия формулируются в виде следующих *постулатов*.

1. Случайная величина Y и ее ошибка ε подчинены нормальному закону распределения.
2. Дисперсия выходной величины Y постоянна и не зависит от величины Y_i, i =1,2,...n.
3. Результаты наблюдений Y_i в разных точках эксперимента независимы и не коррелированны.

К этим постулатам добавляют еще один, который практически в большой степени обеспечивает выполнение первых трех.

4. Входные переменные X_j - независимы, неслучайны, измеряются без ошибок.