Проверка адекватности линейной регрессии

Определение: **Адекватность регрессионного уравнения**, это соответствие его реальному моделируемому процессу, достоверность его параметров.

Схема проверки адекватности уравнения

1. Анализируются показатели качества подгонки регрессионного уравнения ;
2. Проверяются различные гипотезы относительно параметров регрессионного уравнения ;
3. Проверяется выполнение условий для получения «достоверных» оценок методом наименьших квадратов;
4. Производится содержательный анализ регрессионного уравнения.

Проверка качества подгонки

Показатели качества подгонки отражают соответствие расчетных значений зависимой переменной $\hat{y}$ фактическим значениям зависимой переменной y. Эти показатели основываются на $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Первый показатель — **остаточная дисперсия**. Для однофакторного уравнения остаточная дисперсия вычисляется по формуле : $\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$

Чем меньше σ^2, тем лучше регрессионное уравнение описывает моделируемый процесс. σ^2 является размерной величиной и сопоставление регрессионных уравнений, отражающих различные переменные, измеренные в различных единицах измерения, невозможно.

Второй показатель — **коэффициент детерминации R^2**.

Коэффициент детерминации вычисляется по формуле : $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Коэффициент детерминации принимает значения в интервале от 0 до 1. Чем ближе R^2 к единице, тем лучше качество подгонки регрессионного уравнения, так как R^2 приближается к единице при приближении вычитаемой дроби к 0. В свою очередь указанная дробь приближается к нулю при приближении к нулю числителя, то есть при небольших отклонениях фактических и теоретических значений зависимой переменной. На основании R^2 возможно сопоставление различных уравнений.

Проверка гипотеза о том, что линейная связь между *x* и *y* не подтверждается

Отсутствие связи можно изучить на основе отклонений расчетных значений $\hat{y}_i$ от среднего арифметического значения $\bar{y}$ и отклонения расчетных значений $\hat{y}_i$ от фактических значений y_i Близкое к нулю значение $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ свидетельствует об отсутствии какой-либо тенденции для y_i в связи с изменением x.

H_0: $\hat{a} = \hat{b} = 0$, (т.е. линейная связь между *x* и *y* отсутствует);

H_1: $\hat{a}^2 + \hat{b}^2 \neq 0$, (т.е. наличие линейной связи).

Рассчитываем значение *F*-статистики $F_{расч} = \dfrac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2 / (n-2)} = \dfrac{\sum(\hat{y}_i - \bar{y})^2}{\sigma^2}$

$F_{табл} = F^{p}_{1,\ n-2}$ - табличное значение распределения Фишера для вероятности *p* и степеней свободы $m_1=1$, $m_2=n-2$.

$F_{табл} > F_{расч} \Rightarrow$ принимаем H_0 с вероятностью *p;*

$F_{табл} < F_{расч} \Rightarrow$ отвергаем H_0 в пользу H_1 с вероятностью *p*.

Проверка гипотез относительно параметров регрессионного уравнения

Отдельно исследуется коэффициент регрессии ***b***. Выдвигается гипотеза о том, что ***x*** влияет на ***y*** несущественно, то есть ***y*** изменяется по каким-то другим причинам, а не в связи с изменениями ***x***.

H_0: $\hat{b} = 0$, (т.е. фактор *x* незначим);

H_1: $\hat{b} \neq 0$, (т.е. фактор *x* значим).

t-статистика считается по формуле: $t = \frac{\hat{b} - b}{\sigma_b} = \frac{\hat{b}}{\sigma_b}$

где σ_b — стандартная ошибка коэффициента *b*, вычисляемая по формуле:

$$\sigma_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

По общей процедуре проверки гипотез находим $t^{табл}$ (в таблице Стьюдента) с заданным уровнем значимости α (вероятностью p=1-α) и степенями свободы v=n-2.

Если $|t^{расч}| > t^{табл}$, то с заданной вероятностью гипотезу b=0 отвергаем.

Аналогично проверяется гипотеза о значимости свободного члена *a* в уравнении регрессии.

Проверка достоверности оцененных параметров регрессионного уравнения

Возможность применения регрессионного уравнения определяются достоверностью оцененных параметров модели или, по другому, «хорошими» свойствами оценок коэффициентов регрессии: несмещенностью, состоятельностью и эффективностью оценок.

Параметры регрессионного уравнения, полученные методом наименьших квадратов, являются достоверными тогда и только тогда, когда остаточная компонента ε уравнения удовлетворяет условиям:

1. Остаточная компонента носит случайный характер.
2. $M(\varepsilon_i) = 0$ -мат. ожидание случайной компоненты равно нулю,
3. $D(\varepsilon_i) = \sigma^2 = const$ - дисперсия случайной компоненты — постоянна,
4. $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$ - отсутствует автокорреляция;
5. $\varepsilon_i \sim N(0, \sigma^2)$ Нормальность распределения.

Проверка случайности остаточной компоненты

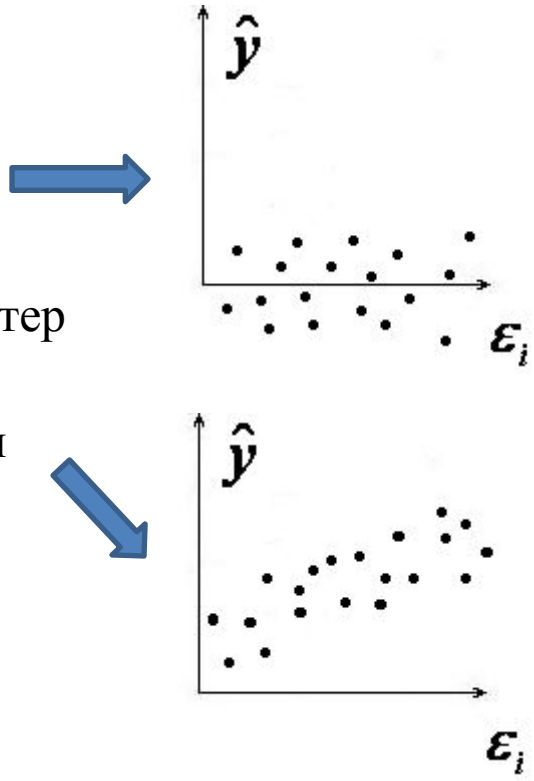
Для проверки случайного характера остатков ε строят график зависимости остатков ε_i от расчетных значений зависимой переменной $\hat{y}$.

Если на графике нет направленности в расположении точек ε_i, то остатки ε случайные величины.

Если ε зависит от $\hat{y}$, то остаточная компонента ε не случайна.

Остатки – носят систематический характер

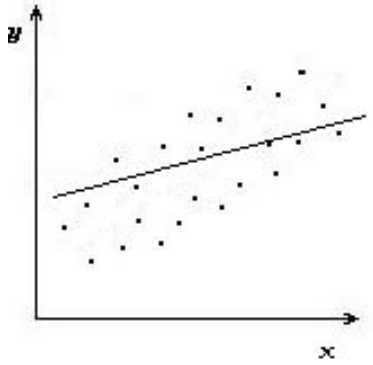
В этих случаях возможно следовало выбрать в качестве регрессионной связи нелинейную зависимость.



Выполнение предпосылки МНК

Проверка условия $M(\varepsilon_i) = 0$

Выполнение этой предпосылки означает получение несмещенных оценок.



$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

В случае, когда значение $M(\varepsilon_i) = 0$, для проверки соответствующей предпосылки применяю следующий тест:

H_0: $M(\varepsilon_i) = 0$, (математическое ожидание остатков равно нулю);

H_1: $M(\varepsilon_i) \neq 0$, (математическое ожидание остатков отлично от нуля).

Рассчитывается значение критерия $t_{расч} = \dfrac{\mu}{\sigma / \sqrt{n}}$

где $\sigma = \sqrt{\dfrac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-1}}$ - несмещенное выборочное стандартное отклонение, μ - выборочное среднее. $t_{табл} = t^{p}_{n-1}$ - табличное значение распределения Стьюдента для вероятности p и степени свободы $m=n-1$.

$|t_{расч}| < t_{табл} \Rightarrow$ принимаем H_0 с вероятностью p;

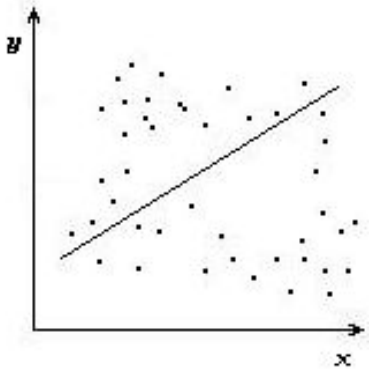
$|t_{расч}| > t_{табл} \Rightarrow$ отвергаем H_0 в пользу H_1 с вероятностью p.

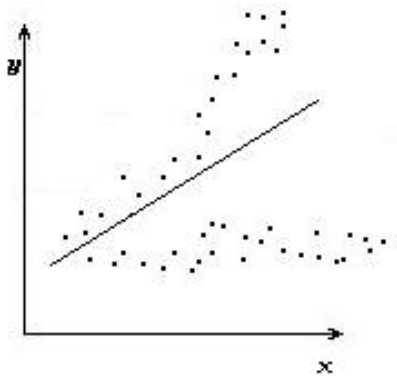
Выполнение предпосылки МНК

Проверка условия $D(\varepsilon_i) = \sigma^2 = const$

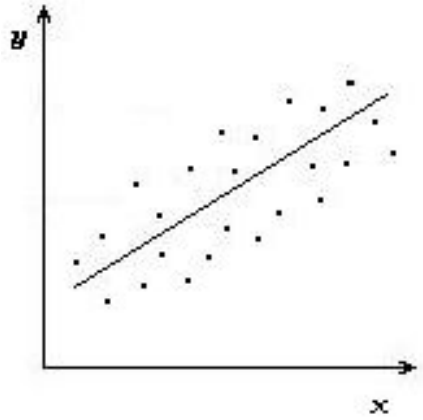
Выполнение этой предпосылки означает получение эффективных оценок.

Определение. Выполнение условия постоянства дисперсии (отсутствие ее роста с ростом независимой переменной) называется **гомоскедастичностью**. В противном случае **гетероскедастичностью**.





$D(\varepsilon_i) \neq \sigma^2$ -гетероскедастичность



$D(\varepsilon_i) = \sigma^2$

кедастичность

Проверка выполнения условия о постоянстве дисперсии остатков

Тест Гольфелда-Квандта

1. упорядочение n наблюдений по мере возрастания переменной x;
2. исключение из рассмотрения C центральных наблюдений, при этом $(n\text{-}C)/2>p$, где p- число оцениваемых параметров;
3. разделение совокупности из $(n\text{-}C)$ наблюдений на две группы (соответственно с малыми и большими значениями фактора x) и определение по каждой из групп уравнений регрессий;
4. определение остаточной суммы квадратов для первой (S_1) и второй(S_2) групп и нахождение их отношения $R = \frac{S_1}{S_2}$, где $S_1 > S_2$.

При выполнении нулевой гипотезы о гомоскедастичности остатков отношение R будет удовлетворять F-критерию с $(n\text{-}C\text{-}2p)/2$ степенями свободы для каждой остаточной суммы квадратов. Чем больше величина R превышает табличное значение F-критерия, тем более нарушена предпосылка о равенстве дисперсий остаточных величин.

Проверка выполнения условия о постоянстве дисперсии остатков

Применение теста Гольфелда-Квандта (схема)

1. Все n наблюдений упорядочиваются по величине x_j.
2. Вся упорядоченная выборка разбивается на три подвыборки: определяем количество отбрасываемых наблюдений из расчета $n\backslash 6$.
3. Оцениваются отдельные регрессии для первой подвыборки (k первых наблюдений) и для третьей подвыборки (k последних наблюдений).
4. Определить остатки (ошибки) для первой и последней группы.
5. Возводим каждую группу остатков в квадрат и суммируем их.
6. Сравниваем две полученные суммы при этом разделим наибольшую из них на наименьшую (это будет $F_{расч}$).
7. Определяем $F_{табличное}$ со степенями свободы $n_1=n_1\text{-}2$ и $n_2= n_2\text{-}2$, где $n_{1,2}$- количество наблюдений в первой и соответственно во второй группе
8. Сравнить $F_{расч}$ с $F_{табл}$. Если первое меньше второго, то есть рост дисперсии с увеличением независимого фактора (имеется гетероскедостичность) и наоборот.

Проверка выполнения условия о постоянстве дисперсии остатков

Тест Спирмена.

Суть теста заключается в определении наличия связи между ростом остаточной компоненты и ростом независимого фактора, то есть определение роста дисперсии остатков. Проверяется такая зависимость на основе расчета коэффициента ранговой корреляции Спирмена ρ между остатками модели ε и независимым фактором x. Проверка статистической значимости коэффициента Спирмена на основе соответствующего t-критерия аналогична проверке нулевой гипотезы об отсутствии гетероскедастичности в остатках.

Существуют и другие тесты для определения гетероскедастичности в остатках, например тест Глейзера, Уайта.

Схема теста Спирмена

1. Проранжировать значения независимой переменной X, присваивая ранг 1 наименьшему значению, и т.д. Занести ранги в первый столбец таблицы по порядку номеров испытуемых или признаков.
2. Проранжировать значения ряда остатков ε, в соответствии с теми же правилами. Занести ранги во второй столбец таблицы по порядку номеров испытуемых или признаков.
3. Подсчитать разности d между рангами X и ε по каждой строке таблицы и занести в третий столбец таблицы.
4. Возвести каждую разность в квадрат: d^2. Эти значения занести в четвертый столбец таблицы.
5. Подсчитать сумму d^2.
6. При наличии одинаковых рангов рассчитать поправки:

где a - объем каждой группы одинаковых рангов в ранговом ряду X; b - объем каждой группы одинаковых рангов в ранговом ряду ε.

$$T_a = \sum \left(a^3 - a\right)/12$$

$$T_b = \sum \left(b^3 - b\right)/12$$

Схема теста Спирмена

8. Рассчитать **коэффициент ранговой корреляции** r_s по формуле:

при отсутствии одинаковых рангов

$$r_s = 1 - 6 \cdot \frac{\sum d^2}{N \cdot \left(N^2 - 1\right)}$$

при наличии одинаковых рангов

$$r_s = 1 - 6 \cdot \frac{\sum d^2 + T_a + T_b}{N \cdot \left(N^2 - 1\right)}$$

где sum(d^2) - сумма квадратов разностей между рангами;

T_a и T_b - поправки на одинаковые ранги;

N - количество наблюдений признаков, участвовавших в ранжировании.

Проверка значимости коэффициента ранговой корреляции Спирмена

Нулевая и альтернативная гипотезы имеют вид:

H_0*: *коэффициент ранговой корреляции Спирмена r_s незначимый, гетероскедастичности нет;

H_1*: *коэффициент ранговой корреляции Спирмена r_s значим, гетероскедастичность есть

- Рассчитывается t-статистика по формуле:

$$t_{расч.} = \frac{r_s}{\sqrt{1 - r_s^2}} \sqrt{(n-2)}$$

- Определяется $t_{табл}$ по таблице Стьюдента со степенями свободы n-2 и уровнем значимости α
- *Если* $|t_{расч}| > t_{табл}$, *то* $\boldsymbol{H_0}$ отклоняют на заданном уровне значимости, и считаем, что имеет место гетероскедастичность остатков.

Проверка выполнения условия о постоянстве дисперсии остатков

Определение: нарушение условия независимости между ошибками для разных наблюдений называется **автокорреляцией** в остатках. То есть имеется зависимость случайных компонент для наблюдений с различными номерами (i и j).

Нарушение условия $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$ приводит к получению неэффективных оценок и как следствие невозможности применения полученных моделей в прогнозных целей, в силу ненадежности полученных результатов.

Автокорреляцию можно представить в виде авторегрессии различного порядка, так, например, если текущее значение остатков ε_i находится в линейной зависимости от предыдущего порядка ε_{i-1}($\varepsilon_i = \rho\varepsilon_{i-1} + e_i$, (*)), то имеет место **авторегрессия первого порядка (AR(1)),** если имеет место влияние предпредыдущих значений остатков ε_{i-1} , то есть

$\varepsilon_i = \rho_1\varepsilon_{i-1} + \rho_2\varepsilon_{i-2} + e_i$, то имеет место **авторегрессия второго порядка (AR(2))**.

Считаем, что номера наблюдений упорядочены по возрастанию номера наблюдения i.

Тест на определение автокорреляции в остатках

Тест Дарбина-Уотсона

Тест Дарбина-Уотсона: обнаружение автокорреляции остатков вида $\varepsilon_i = \rho\varepsilon_{i-1} + e_i$

То есть представленных в виде авторегрессии первого порядка. .

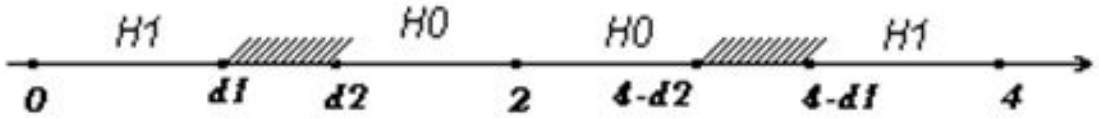
H_0: $\rho = 0$, (т.е. автокорреляция остатков отсутствует);

H_1: $\rho > 0$ или $\rho < 0$, (наличие положительной или отрицательной автокорреляции остатков).

Расчетное значение статистики Дарбина-Уотсона:

$$dw = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^{\,2}}$$

d_1, d_2- табличные значения распределения Дарбина-Уотсона для степеней свободы *n,* и вероятности *p.* Области принятия соответствующих гипотез:



$d_1 < dw < d_2$ и $4 - d_2 < dw < 4 - d_1$ - зона неопределенности

При проверке наличия автокорреляции на практике руководствуются простым правилом: расчетное значение *D-W*, близкое к 2, свидетельствует об отсутствии автокорреляции. Значение близкое к 4 свидетельствует об отрицательной автокорреляции, а близкое к нулю — о положительной.

Наличие авторегресии II порядка проверяют с тестом Броша-Годфри.

Тест Бройша- Годфри

Тест Дарбина-Уотсона нельзя применять в случае, если:

1. В модели содержаться лаговые переменные (сдвинутые на определенный временной интервал вперед или назад)
2. В модели есть автокорреляция, выраженная авторегрессией второго и более высоких порядков.
3. В модели нет свободного члена
4. Количество наблюдений, по которым строилась модель, достаточно мало.

Тест Бройша- Годфри: рассматривается $\varepsilon_i = \sum_{j=1}^{k} \alpha_j \cdot \varepsilon_{i-j} + u_i$

H_0: $\hat{\alpha}_1 = \hat{\alpha}_2 = \ldots = \hat{\alpha}_k$ (автокорреляция, выраженная авторегрессией *k-ого* порядка, отсутствует);

H_1: $\hat{\alpha}_1^2 + \hat{\alpha}_2^2 + \ldots + \hat{\alpha}_k^2 \neq 0$ (автокорреляция в остатках имеется).

Рассчитывается $LR=nR^2$ статистика подчиняется χ^2-распределению с k степенями свободы. Здесь R^2 –коэффициент детерминации, n – общее число кросс-секций. Если табличное значение $\chi^2<LR$, то автокорреляции в остатках нет.

Условие нормальности остатков

Нарушение условия $\varepsilon_i \sim N(0,\sigma^2)$ приводит к получению несостоятельных оценок, и как следствие приводящих к ненадежным прогнозам.

Критерий Колмогорова-Смирнова

H_0: $F(\varepsilon) = F_0(\varepsilon)$, $F_0(\varepsilon)$ где - функция нормального распределения (распределение остатков согласуется с нормальным распределением);

H_1: $F(\varepsilon) \neq F_0(\varepsilon)$, (распределение остатков не согласуется с нормальным распределением).

$KS_{\text{табл}} > KS_{\text{расч}} \Rightarrow$ принимаем H_0 с вероятностью p;

$KS_{\text{табл}} < KS_{\text{расч}} \Rightarrow$ отвергаем H_0 в пользу H_1 с вероятностью p.

Тест Бера-Жарка

Соответствие распределения остатков модели нормальному закону можно проверить с помощью теста Бера-Жарка, для которого определяется *JB*-статистика по формуле:

$$JB = (n-k-1)\cdot\left(\frac{m_1^2}{6} - \frac{(m_2-3)^2}{24}\right)$$

где $m_1 = \left(\frac{\sum_i (\varepsilon_i - \bar{\varepsilon})^3}{n}\right) \Big/ \sqrt{\left(\frac{\sum_i (\varepsilon_i - \bar{\varepsilon})^2}{n}\right)^3}$ – коэффициент асимметрии распределения остатков, $m_2 = \left(\frac{\sum_i (\varepsilon_i - \bar{\varepsilon})^4}{n}\right) \Big/ \sqrt{\left(\frac{\sum_i (\varepsilon_i - \bar{\varepsilon})^2}{n-1}\right)^4}$ – коэффициент эксцесса, n – объем выборки, $\bar{\varepsilon}$ – среднее значение остатков, k – количество независимых факторов в модели.

Нулевая гипотеза о «ненормальности» распределения остатков отклоняется на выбранном уровне значимости, если $JB > \chi^2_{табл}$, определённого для степеней свободы $n–p–q$ из таблицы критических значений χ^2-распределения.

Применение регрессионных уравнений

Пример. Зависимость урожайности какой-то культуры от уровня внесения удобрений описывается следующей регрессионной моделью $y = 11{,}2 + 6{,}23 \cdot x$, где y — то урожайность, а x уровень внесения удобрений.

Определим $\tilde{y}_1$ при $\tilde{x}_1 = 1{,}0$ ц.: $\tilde{y}_1 = 11{,}2 + 6{,}23 \cdot 1 = 17{,}43$. Затем увеличим внесение удобрений на 1,0 ц., то есть $x_2 = 2{,}0$. Тогда $\tilde{y}_2 = 11{,}2 + 6{,}23 \cdot 2 = 23{,}66$

Найдем $\Delta y = y_2 - y_1 = 6{,}23$

Следовательно, коэффициент регрессии показывает прирост зависимой переменной приходящийся на единицу прироста независимой переменной. Коэффициент регрессии является размерной величиной и абсолютная величина его зависит от единиц измерения x и y. В нашем случае единица измерения коэффициента регрессии ц/ц.