МНОЖЕСТВЕННАЯ

РЕГРЕССИЯ

в эконометрических расчетах

ВАЖНОСТЬ ИСПОЛЬЗОВАНИЯ

- При использовании парной регрессии предполагается, что влиянием других факторов на результат можно пренебречь (сделать их неизменными)
- В реальной практике экономические данные зафиксировать не удается и чистое влияние двух переменных друг на друга выделить нельзя, поэтому используется множественная регрессия, дополнительные факторы вводят в модель

СФЕРА ПРИМЕНЕНИЯ

Решение задач оценки

- объема спроса,
- доходности акций
- плановых издержек
- макроэкономических прогнозов

Цель применения

- Построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности на результат, а также совокупное воздействие их на моделируемый показатель

Основные предпосылки модели множественной регрессии

- Математическое ожидание всех ε_i равно нулю для всех наблюдений;
- Дисперсии всех ε_i постоянны и равны;
- ε_i – независимы друг от друга и от $x_1 \ldots x_p$;
- ε_i – имеют распределение Гаусса $N(0;\sigma^2)$;
- Модель линейна относительно параметров $ß_1 \ldots ß_p$;
- Между $x_1 \ldots x_p$ отсутствует строгая линейная связь (нет мультиколлинеарности факторов);

Предпосылки МНК (условия Гаусса–Маркова)

1 . Математическое ожидание случайного отклонения ε_i равно нулю: $M(\varepsilon_i) = 0$ для всех наблюдений.

Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную. В каждом конкретном наблюдении случайный член может быть либо положительным, либо отрицательным, но он не должен иметь систематического смещения.

2 . Дисперсия случайных отклонений ε_i постоянна:

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2 \text{ для любых наблюдений } i \text{ и } j.$$

Данное условие подразумевает, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть либо большим, либо меньшим, не должно быть некой априорной причины, вызывающей большую ошибку (отклонение).

Выполнимость данной предпосылки называется *гомоскедастичностью (постоянством дисперсии отклонений).* Невыполнимость данной предпосылки называется *гетероскедастичностью (непостоянством дисперсий отклонений).*

3 . Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$.

Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями. Другими словами, величина и определенный знак любого случайного отклонения не должны быть причинами величины и знака любого другого отклонения.

если данное условие выполняется, то говорят об отсутствии *автокорреляции.*

4 . Случайное отклонение должно быть независимо от объясняющих переменных.

Обычно это условие выполняется автоматически при условии, что объясняющие переменные не являются случайными в данной модели.

Следует отметить, что выполнимость данной предпосылки не столь критична для эконометрических моделей.

5 . Модель является линейной относительно параметров.

Теорема Гаусса–Маркова. Если предпосылки *1 – 5* выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются несмещенными, т. е. $M(b_0) = в_0, M(b_1) = в_1$. Это вытекает из того, что $M(e_i) = 0$ и говорит об отсутствии систематической ошибки в определении положения линии регрессии.
2. Оценки состоятельны, т. к. дисперсия оценок параметров при возрастании числа n наблюдений стремится к нулю: $D(b_0) \xrightarrow[n\to\infty]{} 0$, $D(b_1) \xrightarrow[n\to\infty]{} 0$. Другими словами, при увеличении объема выборки надежность оценок увеличивается (b_0 наверняка близко к β_0, b_1 – близко к β_1).
3. Оценки эффективны, т. е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин y_i.

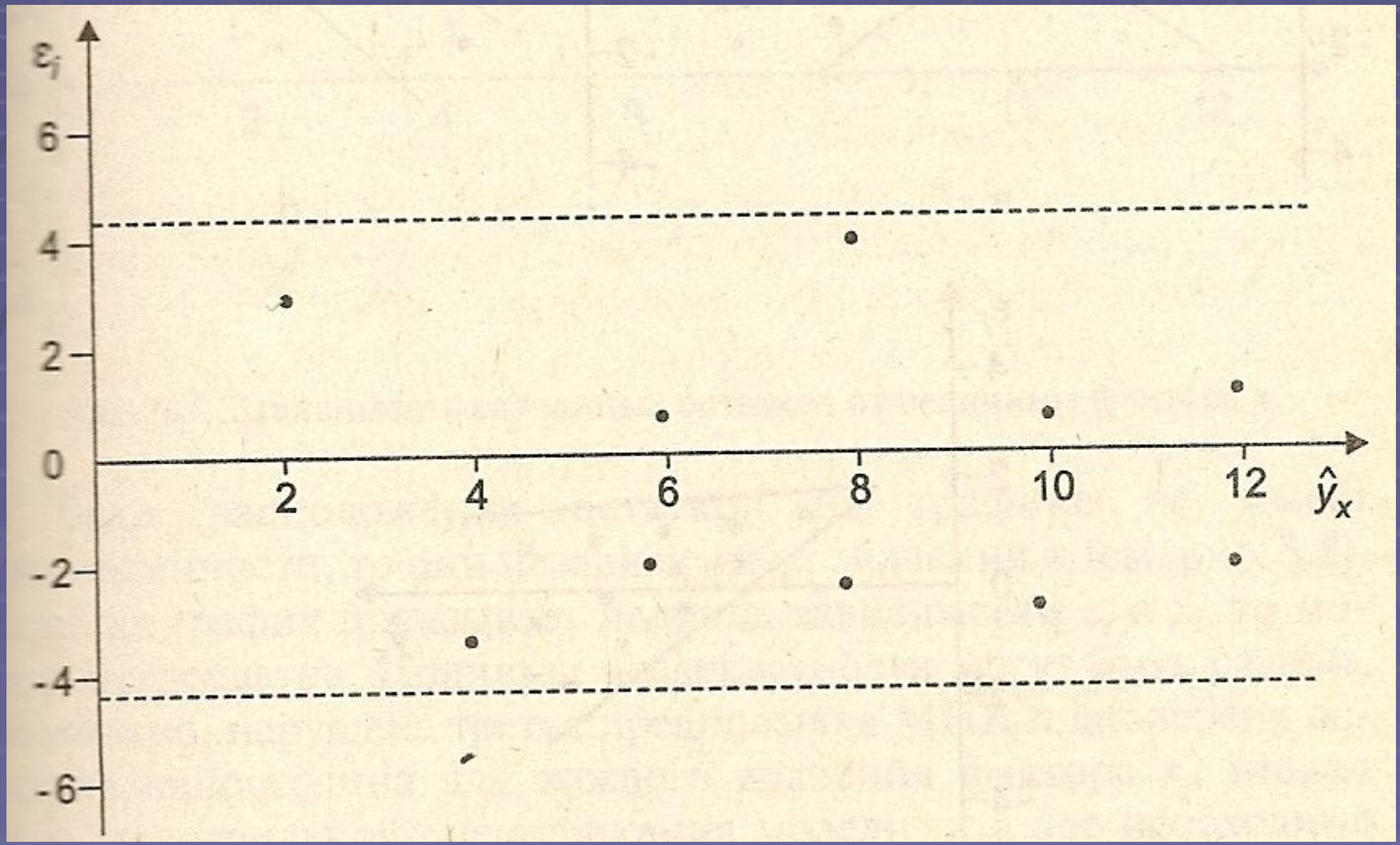
В англоязычной литературе такие оценки называются *BLUE (Best Linear Unbiased Estimators) – наилучшие линейные несмещенные оценки.*

Если предпосылки *2* и *3* нарушены, т. е. дисперсия отклонений непостоянна и (или) значения e_i, e_j связаны друг с другом, то свойства несмещенности и состоятельности сохраняются, но свойство эффективности – нет.

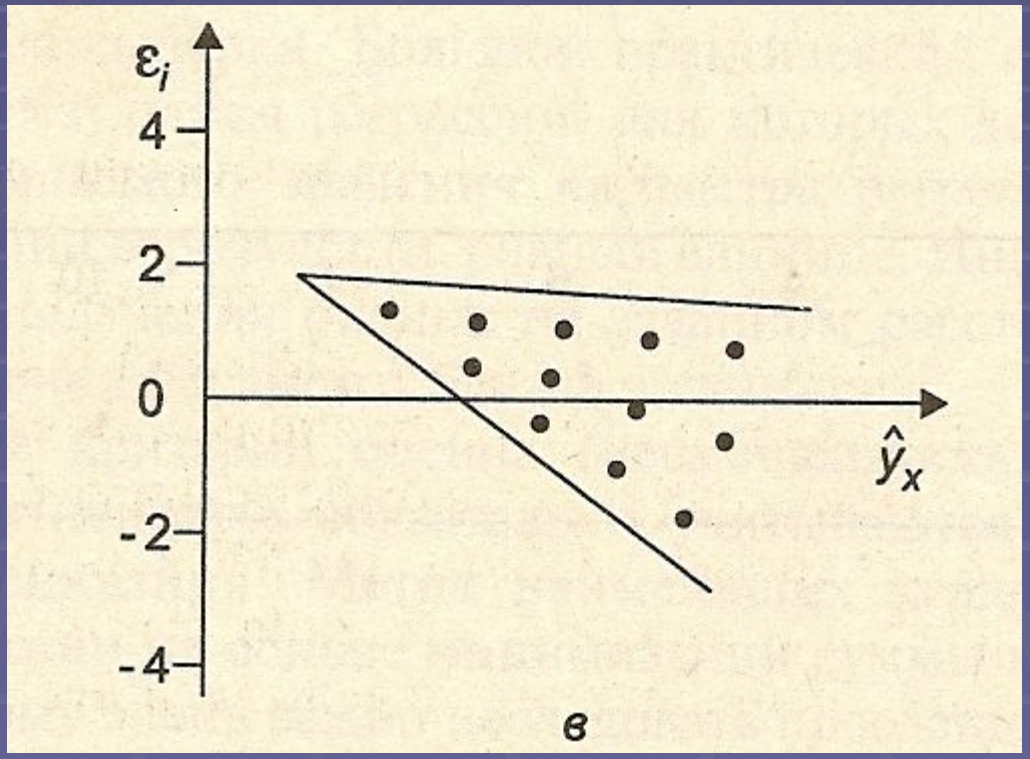
Наряду с выполнимостью указанных предпосылок при построении классических линейных регрессионных моделей делаются еще некоторые предположения :

- объясняющие переменные не являются случайными величинами;
- случайные отклонения имеют нормальное распределение;
- число наблюдений существенно больше числа объясняющих переменных;
- отсутствуют ошибки спецификации;
- отсутствует мультиколлинеарность.

Остатки случайны

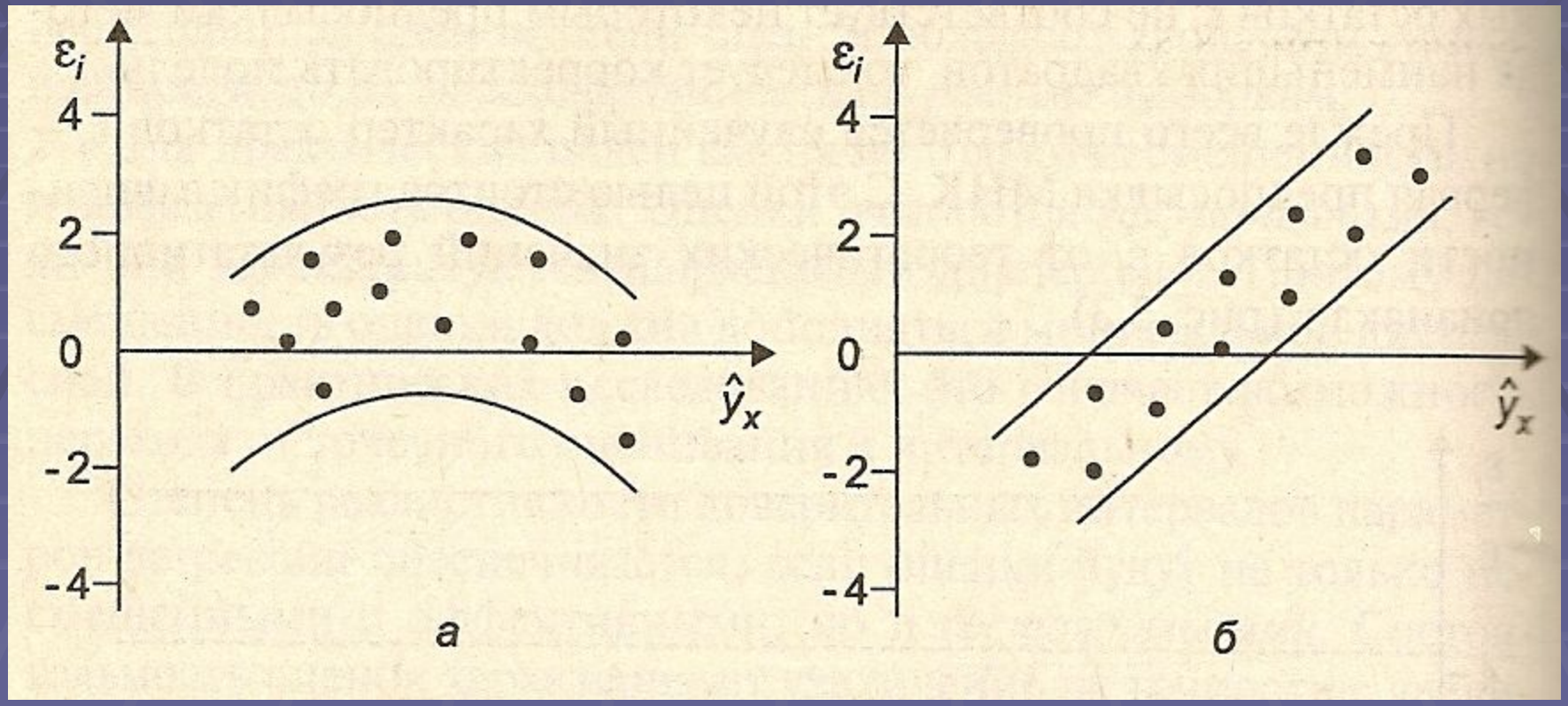


Графический анализ остатков



- Остатки не являются случайными величинами

Графический анализ остатков



- Остатки не являются случайными величинами

НУЖНО

- Применить другую функцию

или

- Добавить информации , пока остатки не станут случайными

УРАВНЕНИЕ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$\hat{Y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p, \quad m = p + 1$$

По МНК вектор оценок параметров модели регрессии находится по формуле:

$$a = (X'X)^{-1} X'Y$$

Значимость уравнения подтверждается коэффициентом детерминации

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

Критерий значимости Фишера, n – число наблюдений, m – число параметров в модели регрессии (число коэффициентов регрессии):

$$F = \frac{R^2 (n - m)}{(1 - R^2)(m - 1)} > F_\alpha (k_1 = m - 1; k_2 = n - m)$$

МАТРИЦЫ X ,У, A и E

$$X = \begin{pmatrix} 1 & x_{11} & \boxtimes & x_{p1} \\ 1 & x_{12} & \boxtimes & x_{p2} \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ 1 & x_{1n} & \boxtimes & x_{pn} \end{pmatrix}, \text{ поэтому } m = p+1$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \boxtimes \\ y_n \end{pmatrix} \quad A = \begin{pmatrix} a_0 \\ a_2 \\ \boxtimes \\ a_p \end{pmatrix} \quad E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \boxtimes \\ \varepsilon_n \end{pmatrix}$$

СКОРРЕКТИРОВАННЫЙ R^2

Чтобы получить более объективную оценку качества уравнения регрессии R^2 корректируют на количество наблюдений и факторов

$$R^2_{ck} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Доверительные интервалы для среднего значения Y и индивидуального значения Yi в случае множественной регрессии

$$\hat{Y}_i \pm t_{табл}(n-m)S_{\hat{Y}_i}, где\ S_{\hat{Y}_i} = S\sqrt{X_i(X^T X)^{-1}X_t^T},$$

$$S = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-p-1}}, S^2 - остаточная\ дисперсия$$

$$\hat{Y}_i \pm t_{табл}(n-m)S_{\hat{Y}_i}, где\ S_{\hat{Y}_i} = S\sqrt{1 + X_i(X^T X)^{-1}X_t^T}$$

МЕТОДЫ ПОСТРОЕНИЯ УРАВНЕНИЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

- Метод исключения (отсев фактора из полного набора)
- Метод включения (введение нового фактора)
- Шаговый регрессионный анализ (исключение введенного ранее фактора)

ОТСЕВ ФАКТОРОВ

- 1 путь. Проводится по показателям не парной , а частной корреляции, которые в чистом виде оценивают взаимосвязь между фактором и результатом. Строится матрица частных коэффициентов корреляции
- 2 путь. По критерию Стьюдента из уравнения исключаются те факторы, у которых значение критерия меньше табличного

ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Позволяет установить степень «чистого» влияния факторного признака на результативный признак, при условии, что остальные факторы не влияют, изменяется от 0 до 1, не может быть больше по величине коэффициента множественной корреляции.

$$r_{yx_k}(x_1, x_2 \ldots x_{k-1}) = \sqrt{\frac{R_k^2 - R_{k-1}^2}{1 - R_k^2}}$$

Где R^2_k – коэффициент множественной детерминации между y и $x_1 \ldots x_к$;

R^2_{k-1} – коэффициент множественной детерминации между y и $x_1 \ldots x_{к-1}$;

ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Если парный коэффициент корреляции между x и y больше частного коэффициента корреляции между x и y, то существует фактор, усиливающий влияние x на y, если наоборот, то существует фактор, ослабляющий это влияние

СПЕЦИФИКАЦИЯ МОДЕЛИ

- Отбор факторов
- Выбор вида уравнения

ОТБОР ФАКТОРОВ

Факторы, включаемые в модель должны удовлетворять требованиям:

1. Быть количественно измеримы или задаваться фиктивными переменными
2. Не должны быть коррелированы между собой (отсутствие мультиколлинеарности)

ОТБОР ФАКТОРОВ

- Включаемые в модель факторы должны объяснять вариацию зависимой переменной
- R^2 – доля объясненной вариации зависимой переменной за счет влияния факторов модели
- $(1-R^2)$ – остаточная дисперсия S^2
- При дополнительном включении в регрессию фактора R^2 должен расти, а S^2 уменьшаться
- Насыщение модели лишними факторами не снижает S^2,но и приводит к статистической незначимости параметров регрессии

ИССЛЕДОВАНИЕ МУЛЬТИКОЛЛИНЕАРНОСТИ

Наличие существенной связи между факторами – мультиколлинеарности факторов - ведет к ненадежности оценок уравнения регрессии и прогнозов на их основе.

Для оценки её наличия используют определитель матрицы парных линейных коэффициентов корреляции между факторами, например для 3 факторов:

$$\det|R| = \begin{vmatrix} r_{x1x1} & r_{x2x1} & r_{x3x1} \\ r_{x1x2} & r_{x2x2} & r_{x3x2} \\ r_{x1x3} & r_{x2x3} & r_{x3x3} \end{vmatrix}$$

Чем ближе значение определителя к нулю, тем сильнее мультиколлинеарность факторов и ненадежней результаты множественной регрессии

Проверка гипотезы о независимости факторов – отсутствии мультиколлинеарности

H_0: Det|R|=1, то есть мультиколлинеарности нет

H_1:Det|R|=0 , то есть она есть

Если $\chi^2_{расч} > \chi^2(\alpha;0,5(m(m-1))$, то H_0 отклоняется и мультиколлинеарность факторов доказана

$X^2_{расч}=[n-1-(1/6)(2m+5)lgDetR]$

УСТРАНЕНИЕ МУЛЬТИКОЛИНЕАРНОСТИ

- Исключение из модели наиболее мультиколлинеарных факторов (строят множественную регрессию относительно каждого фактора и исключают фактор с максимальным R^2)
- Преобразование факторов через их объединение или изменение (Δ)
- Совмещенные уравнения регрессии (при коэффициенте регрессии стоит не один, а произведение факторов)
- Использование уравнений регрессии приведенной формы

РАНЖИРОВАНИЕ ПЕРЕМЕННЫХ X ПО МЕРЕ ИХ ВЛИЯНИЯ НА Y

(если переменные X имеют разные единицы измерения)

на основе коэффициентов эластичности:

$$Э_j = a_j \frac{\bar{x}_j}{\bar{y}}, \quad j = 1..p$$

и стандартизированных коэффициентов регрессии:

$$\beta_j = a_j \frac{s_{x_j}}{s_y}, \quad j = 1..p$$

ПРОВЕРКА ЗНАЧИМОСТИ Xj

На основе t-критерия Стьюдента

$$|t| = \frac{|a_j|}{S_{a_j}} > t_{\alpha/2}(n-m)$$

тогда оценка параметра модели при xj отлична от нуля с вероятностью 1-a

$$S_{a_j} = S\sqrt{\left[(X^T X)^{-1}\right]_{jj}}$$

Измерение системного эффекта на основе уравнения регрессии

- В науке принято изучать влияние не отдельных факторов, а целостные системы факторов и результатов.
- Влияние системы не сводится к арифметической сумме влияний каждого фактора в отдельности, так как возникает «системный эффект» - синергия

Влияние системного эффекта

$$\xi = R^2 - \sum_{j=1}^{p} \beta_j^2$$

ПРИМЕРЫ

Рассмотрим по данным Росстата за 2004 г.[2], включающим основные социально-экономические показатели по субъектам Центрального федерального округа (ЦФО), исключая резко отличающиеся по уровню показателей г. Москву и Московскую область, ряд двухфакторных корреляционных систем связи (см. таблицу 1).

Области	Численность населения, тыс. человек	ОПФ, млрд. рублей	ВРП, млрд. рублей	Занятые в экономике, тыс. человек	Средняя заработная плата, руб./мес.
Белгородская	1512	247	80	668	5294
Брянская	1347	191	49	608	4218
Владимирская	1487	222	67	718	4999
Воронежская	2334	376	105	1066	4570
Ивановская	1115	132	36	476	4087
Калужская	1022	179	53	480	5574
Костромская	718	160	33	328	4784
Курская	1199	249	60	598	4766
Липецкая	1189	257	98	556	5743
Орловская	842	119	46	412	4416
Рязанская	1195	242	66	633	5010
Смоленская	1019	216	52	480	5017
Тамбовская	1145	202	51	518	4108
Тверская	1426	307	74	655	5375
Тульская	1622	267	77	771	5159
Ярославская	1339	359	104	666	6240
Средняя	**1282**	**232,8**	**65,3**	**602**	**4960**
σ	**370,9**	**72,4**	**22,1**	**133,6**	**606,3**

Области	ОПФ на одного занятого, тыс. руб./чел.	Территория, тыс. кв. км	Плотность населения, чел./кв. км
Белгородская	370	27	56
Брянская	317	35	38
Владимирская	309	29	51
Воронежская	353	52	45
Ивановская	277	22	51
Калужская	373	30	34
Костромская	488	60	12
Курская	416	30	40
Липецкая	462	24	50
Орловская	289	25	34
Рязанская	382	40	30
Смоленская	450	50	20
Тамбовская	390	34	34
Тверская	469	84	17
Тульская	346	26	62
Ярославская	539	36	37
Средняя	**389,4**	**37,8**	**38,2**
σ	**75,9**	**16,5**	**14,4**

ПРИМЕРЫ

Система № 1

Y - ВРП (валовой региональный продукт); X_1 - численность населения; X_2 - ОПФ (основные производственные фонды).

Уравнение связи: $\hat{Y} = 1{,}78 + 0{,}00595X_1 + 0{,}2552X_2$,

коэффициент детерминации $R^2_{YX_1X_2} = 0{,}8360$,

стандартизованные коэффициенты регрессии:

$$\beta_1 = 0{,}00595 \times 370{,}9/22{,}1 = 0{,}0999;$$

$$\beta_2 = 0{,}2552 \times 72{,}4/22{,}1 = 0{,}8365.$$

Отсюда: $\beta_1^2 = 0{,}0099$; $\beta_2^2 = 0{,}6989$. Системный эффект ξ есть разность:

$$\xi = R^2 - \sum_{j=1}^{k} \beta_j^2 = 0{,}8360\text{-}0{,}0099\text{-}0{,}6989 = 0{,}1272.$$

Главным фактором, влияющим на вариацию величины валового регионального продукта в областях ЦФО, явилась вариация размеров основных производственных фондов. Вариация численности населения влияет слабо, зато имеется существенный системный эффект совместного влияния размеров ОПФ и численности населения, чем и объясняется достаточно высокий парный коэффициент корреляции ВРП с числом жителей (см. таблицу 2).

Таблица 2

Матрица парных коэффициентов корреляции в системе № 1

Признаки	Y	X_1	X_2
Y	1	0,7325	0,9120
X_1	0,7325	1	0,7563
X_2	0,9120	0,7563	1

МОРАЛЬ

В обыденной жизни эту ситуацию с позиций ВРП можно изложить следующим образом: «У меня есть два друга - один очень близкий, второй мне не очень интересен сам по себе, но он друг моего лучшего друга, и поэтому мы - хорошая компания».

Система № 2

Y - ВРП; X_1 - численность населения; X_2 - территория.

Таблица 3

Матрица парных коэффициентов корреляции в системе № 2

Признаки	Y	X_1	X_2
Y	1	0,7325	0,0363
X_1	0,7325	1	0,1070
X_2	0,0363	0,1070	1

Уравнение связи: $\hat{Y} = 11,87 + 0,0436X_1 - 0,0556X_2$,

при этом коэффициент при X_2 ненадежно отличен от нуля:

$$R^2_{YX_1X_2} = 0,507;$$

$$\beta_1 = 0,0436 \times 370,9/22,1 = 0,7323;$$

$$\beta_2 = -0,0556 \times 11,47/22,1 = 0,0414;$$

$$\beta_1^2 = 0,5362;\ \beta_2^2 = 0,00172.$$

Системный эффект $\xi = 0,507 - 0,5362 - 0,00172 = -0,0309$.

ВЫВОД И МОРАЛЬ

Вариация численности населения существенно влияет на вариацию ВРП, а территории - не влияет существенно. Системный эффект, скорее, отрицателен, но также ненадежен.

На «обыденном языке» это можно изложить так: «У меня есть друг, а второй человек - мне никто. И моего друга он тоже не интересует. Нас троих нельзя считать компанией, хотя и вредит мне этот факт очень мало».

Система № 3

Y - средняя заработная плата; X_1 - фондовооруженность (стоимость ОПФ на одного занятого в экономике); X_2 - плотность населения.

Таблица 4

Матрица парных коэффициентов корреляции в системе № 3

Признаки	Y	X_1	X_2
Y	1	0,6583	0,020
X_1	0,6583	1	-0,509
X_2	0,020	-0,509	1

Уравнение связи: $\hat{Y} = 13,75 + 7,194X_1 + 20,52X_2$.

Оба коэффициента регрессии весьма надежны:

$$R^2_{YX_1X_2} = 0,6013;$$

$$\beta_1 = 0,900; \beta_2 = 0,487; \beta_1^2 = 0,810; \beta_2^2 = 0,238.$$

Оба фактора связаны со средней заработной платой прямой и надежной связью, хотя парный коэффициент корреляции r_{YX_2} несущественно отличен от нуля! Это объясняется обратной связью между плотностью населения и фондовооруженностью. Системный эффект отрицателен: $\xi = 0,6013 - 0,810 - 0,238 = -0,447$. В «переводе» на язык человеческих отношений это выглядит так: «У меня должно было быть два друга, но они в ссоре между собой, и это мне сильно вредит!». И действительно, с точки зрения благосостояния населения, разве хорошо, что в областях с более высокой плотностью населения в среднем наблюдается более низкая фондовооруженность?

Система № 4

Не все возможные сочетания связей удается показать на примере экономики областей ЦФО. Четвертую систему мы иллюстрируем по итогам чемпионата России по футболу за 2005 г. Успехи команд высшей лиги измеряются числом набранных очков - Y. На это число влияют два фактора: число забитых командой мячей (голов) в ворота соперников - X_1 и число пропущенных в свои ворота мячей от соперников - X_2. Исходные данные приведены в таблице 5.

Итоги чемпионата России по футболу, 2005 г.

Команды	Набрано очков	Забито мячей	Пропущено мячей
ЦСКА	62	48	20
Спартак	56	47	26
Локомотив	56	41	18
Рубин	51	45	31
Москва	50	36	26
Зенит	49	45	26
Торпедо	45	37	33
Динамо	38	36	46
Шинник	38	26	31
Томь	37	28	33
Сатурн	33	23	25
Амкар	33	25	36
Ростов	31	26	41
Крылья Советов	29	29	44
Алания	23	27	53
Терек	20	20	50
Средняя	**40,68**	**37,69**	**37,69**
σ	**12,44**	**9,34**	**10,69**

Таблица 6

Матрица парных коэффициентов корреляции в системе № 4

Признаки	Y	X_1	X_2
Y	1	0,9083	-0,8466
X_1	0,9083	1	-0,5989
X_2	-0,8466	-0,5989	1

Уравнение связи: $\hat{Y} = 31{,}4 + 0{,}83385X_1 - 0{,}5584X_2$.

Все коэффициенты высоконадежны: $R^2_{YX_1X_2} = 0{,}9678$;

$$\beta_1 = 0{,}8338 \times 9{,}34/12{,}44 = 0{,}6260;$$

$$\beta_2 = -0{,}5584 \times 10{,}69/12{,}44 = -0{,}4798;$$

$$\beta_1^2 = 0{,}3914;\ \beta_2^2 = 0{,}2302;$$

$$\xi = 0{,}9678 - 0{,}3914 - 0{,}2302 = 0{,}3462.$$

ВЫВОД И МОРАЛЬ

Имеем существенное влияние на вариацию результатов команд обоих факторов и весьма значительный положительный системный эффект, который означает известное футболистам правило: «Хорошее нападение - лучшая защита». Кто больше забивает, тот, в основном, и меньше пропускает. И наоборот, хороший вратарь и защитники - основа успеха нападающих - им не нужно постоянно бегать к своим воротам помогать защите.

На языке личных отношений эта система такова: «У меня есть хороший друг и сильный враг. Но они - враги друг другу и это хорошо для меня! В нашей троице отношения таковы, какими и должны быть!».

Свойства ξ

Системное влияние комплекса факторов на результат может оказаться ослабленным в случае, если различные факторы влияют на результат в разных направлениях, то есть имеется отрицательный «системный эффект». В этом случае влияние конкретного j-го фактора может оказаться более сильным, чем влияние всего комплекса факторов. В силу этого β_j^2 может оказаться по величине больше, чем R^2. Такое положение наблюдается в приведенной выше системе № 3. В некоторых редких случаях β_j^2 может оказаться больше 1. Математического запрета на это не существует. Интерпретировать β_j^2 необходимо как показатель соотношения вариации результата за счет j-го фактора и вариации результата за счет всего комплекса факторов, а не как долю[3], так как показатель доли не может превышать 100%.

Показатель ξ

Показатель системного эффекта является собственно статистическим показателем, поскольку качественное содержание и форма его расчета полностью определяются статистической наукой. По своему содержанию он относится к группе показателей взаимосвязи признаков. Однако он измеряет не силу или тесноту связи, а степень согласованности влияния факторов на результат. Чем больше положительное значение ξ, тем более согласованной является система факторов. Чем больше модуль отрицательного значения ξ, тем менее согласована, более противоречива система факторных признаков.

ВЫБОР ФОРМЫ УРАВНЕНИЯ

- Чаще всего используются линейная и степенная функция
- Чем сложнее функция, тем больше нужно данных
- Использование более сложных уравнений не позволяет осуществить экономическую интерпретацию коэффициентов, это делает их использование менее привлекательным

Смысл коэффициентов линейной модели

- В линейной регрессии свободный член не имеет смысла, коэффициент регрессии означает как в среднем измениться y, если x_i измениться на единицу, а другие факторы будут неизменны

Смысл коэффициентов степенной модели

- Коэффициенты при x являются коэффициентами эластичности и показывают на сколько % измениться y, если x_i измениться на 1% при неизменных других факторах
- Сумма коэффициентов регрессии не всегда равна 1

Гомоскедастичность остатков – предпосылка МНК

- Для каждого x дисперсия остатков одинакова

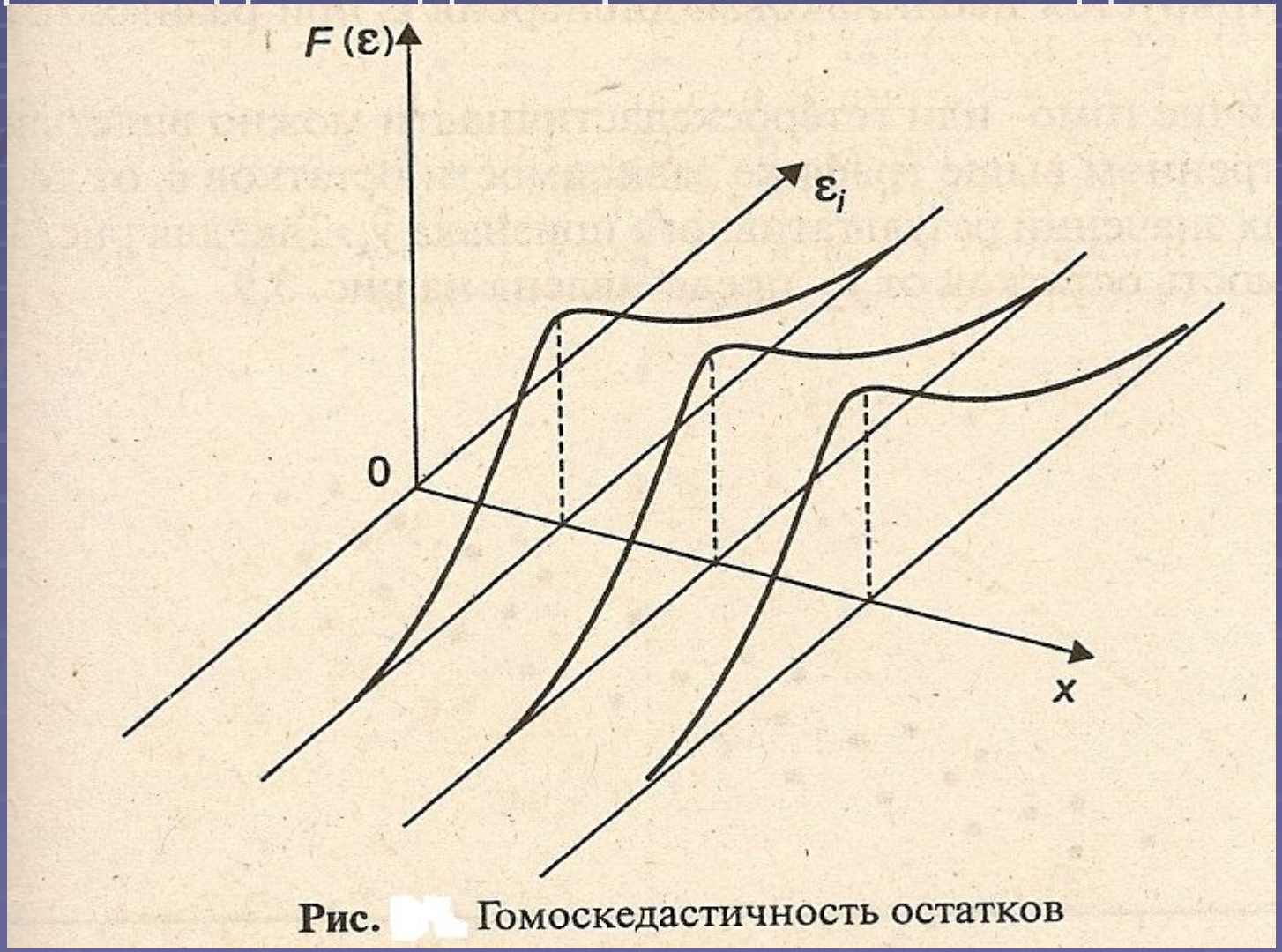


Рис. Гомоскедастичность остатков

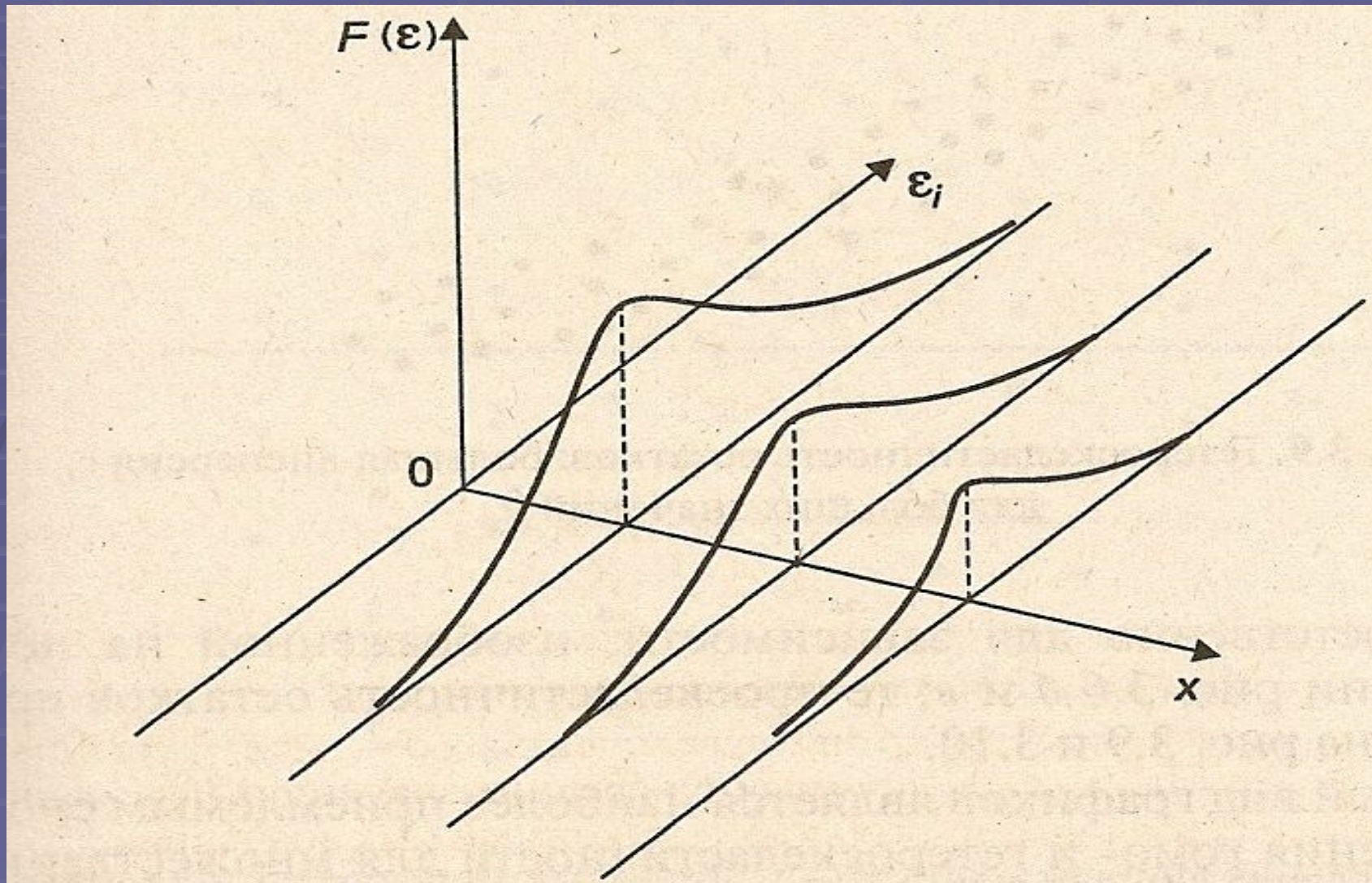


Рис. Гетероскедастичность остатков

Гетероскедастичность остатков

- это **непостоянство дисперсии остатков**, которое также приводит к снижению эффективности применения уравнения регрессии.

Для её выявления используются различные критерии - критерий Голдфелда-Квандта, тест ранговой корреляции Спирмена и д.р.

Тест ранговой корреляции Спирмена

- рассчитывается коэффициент Спирмена между модулями остатков и значениями факторов, если коэффициент Спирмена значим, то гетероскедастичность остатков доказана и уравнение регрессии ненадежно

Тест ранговой корреляции Спирмена

d разность между рангами значений переменной x и $|e_i|$.

Коэффициент ранговой корреляции $r_{x,e} = 1 - 6\dfrac{\sum d^2}{n(n^2-1)}$.

Зададим доверительную вероятность p. $\alpha = (1-p)/2$. По t-таблицам находим граничную точку $t_{\alpha;n-2}$.

Статистика $t = \dfrac{r_{x,e}\sqrt{n-2}}{\sqrt{1-r^2_{x,e}}}$.

Если $t < t_{\alpha;n-2}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности. Иначе гипотеза об отсутствии гетероскедастичности отклоняется. В модели, содержащей несколько факторов, проверка гипотезы об отсутствии гетероскедастичности проводится с помощью статистики t для каждого из них отдельно.

ПРИМЕР

x	e_i	$\lvert e_i \rvert$	d_1	d_2	$d=d_1-d_2$	d^2
2	0	0	5	5	0	0
3	−0,09	0,09	4	2	2	4
4	0,12	0,12	3	1	2	4
5	0,03	0,03	2	4	−2	4
6	−0,06	0,06	1	3	−2	4
Сумма						16

Заполним таблицу. Модули элементов второго столбца запишем в 3-й столбец. В 4-м и 5-м столбцах ранжированы по убыванию элементы 1-го и 3-го столбцов соответственно. $n = 5$ наблюдений.

$$r_{x,e} = 1 - 6\,\frac{\sum d^2}{n(n^2-1)} = 1 - 6\,\frac{16}{5(5^2-1)} = 0{,}2.$$

$\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025$. По t-таблицам находим граничную точку $t_{\alpha;n-2} = t_{0{,}025;5-2} = 3{,}182$.

$$\text{Статистика } t = \frac{r_{x,e}\sqrt{n-2}}{\sqrt{1-r_{x,e}^2}} = \frac{0{,}2\sqrt{5-2}}{\sqrt{1-0{,}2^2}} \approx 0{,}354 < 3{,}182.$$

Мы принимаем гипотезу об отсутствии гетероскедастичности на уровне значимости 5%.

Тест Голдфелда-Квандта

Рассматривается связь величин вида $y = a + bx$. Предполагается, что стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению переменной x в этом наблюдении: $\sigma_i^2 = \sigma^2 x_i^2$, $i = 1, \ldots, n$, n — число наблюдений. Также предполагается, что ε_i имеет нормальное распределение и отсутствует автокорреляция (будет рассмотрена в дальнейшем). Все n наблюдений упорядочиваются по величине x. Эта упорядоченная выборка делится на три примерно равные части объемов k, $n - 2k$ и k соответственно. При $n = 30$ $k = 11$, при $n = 60$ $k = 22$.

Для каждой из выборок объема k оценивается свое уравнение регрессии и находятся суммы квадратов отклонений $S_1 = \sum_{i=1}^{k} e_i^2$ и $S_3 = \sum_{i=n-k+1}^{n} e_i^2$ соответственно.

Зададим доверительную вероятность p. $\alpha = 1 - p$. По F-таблицам находим граничную точку $F_{\alpha;k-m-1;k-m-1}$, где m — число факторов модели.

Статистика $F = S_3/S_1$.

Если $F < F_{\alpha;k-m-1;k-m-1}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности. Иначе гипотеза об отсутствии гетероскедастичности отклоняется. Для множественной регрессии тест обычно проводится для того фактора, который в максимальной степени связан

с σ_i. При этом выбирают $k > m + 1$. Если нет уверенности относительно выбора фактора x_j, то данный тест можно осуществить для каждого фактора.

Пример 34. Рассматривается регрессионная линейная модель с $m = 2$ факторами. $n = 30$ наблюдений. Для первых и последних $k = 11$ наблюдений суммы квадратов отклонений $S_1 = 20$ и $S_3 = 45$ соответственно. С помощью теста Голдфелда-Квандта проверим гипотезу об отсутствии гетероскедастичности.

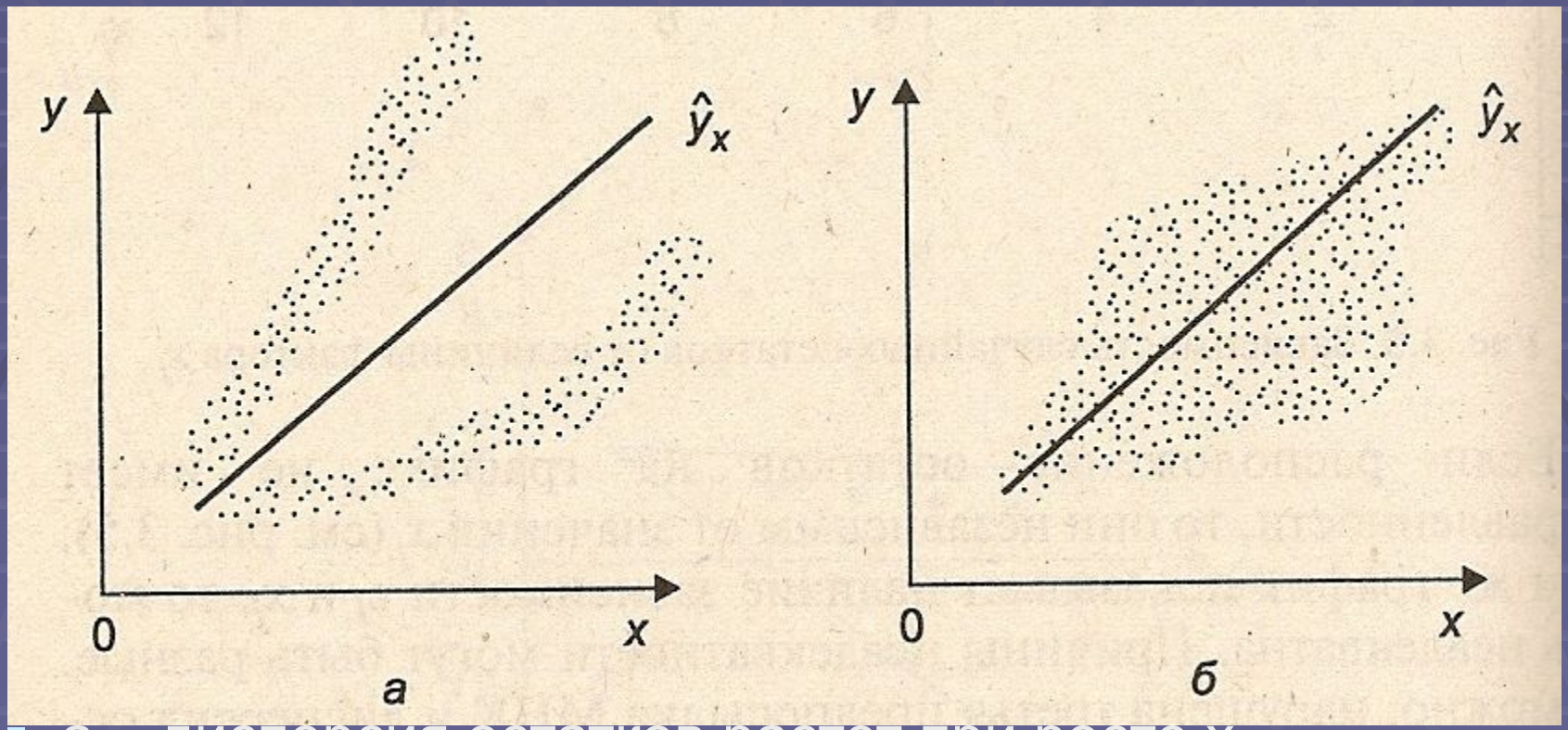
Зададим доверительную вероятность $p = 95\%$.

$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. По F-таблицам находим граничную точку $F_{\alpha;k-m-1;k-m-1} = F_{0{,}05;11-2-1;11-2-1} = 3{,}44$.

Статистика $F = S_3/S_1 = 45/20 = 2{,}25 < 3{,}44$.

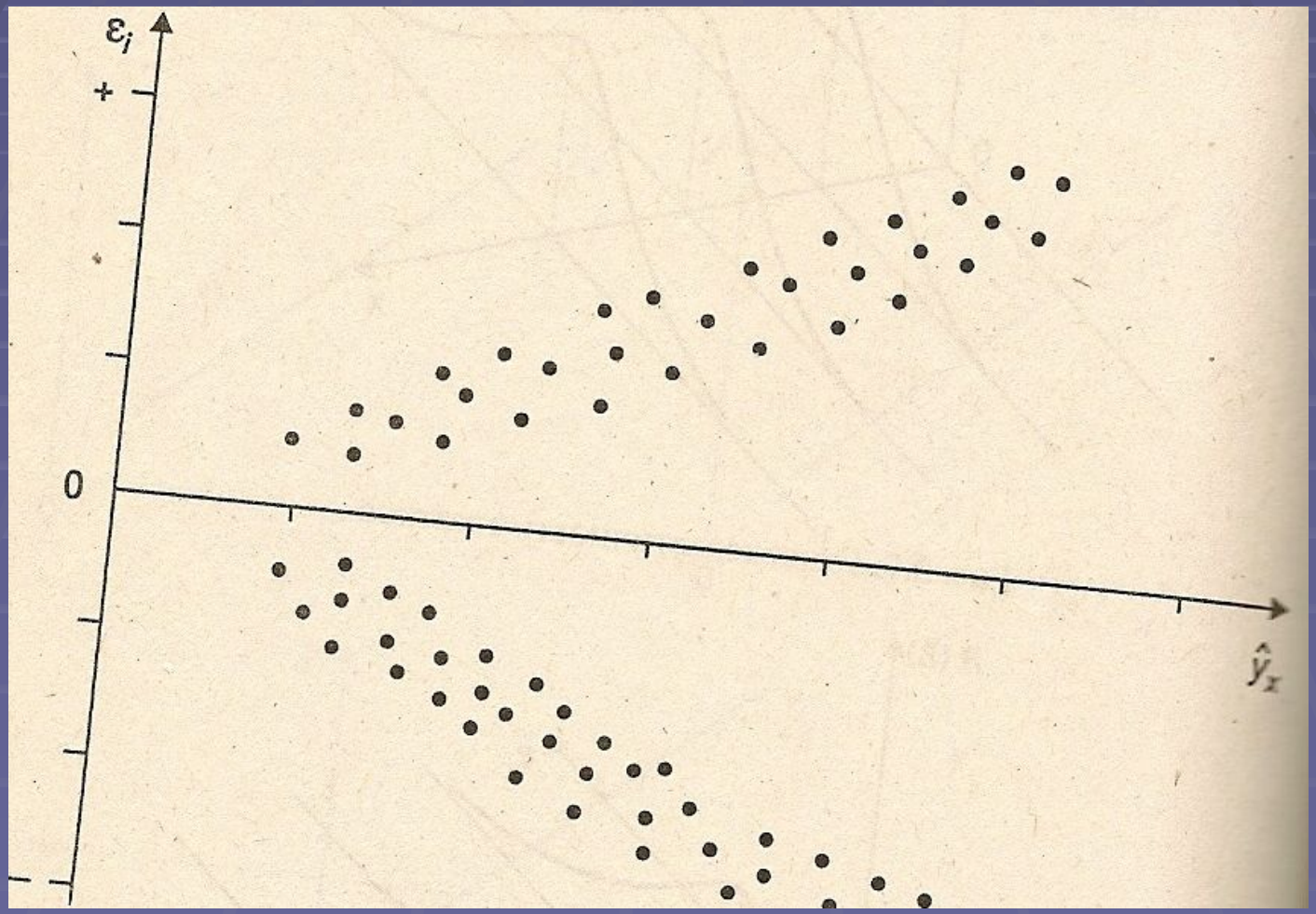
На уровне значимости 5% принимается гипотеза об отсутствии гетероскедастичности.

Графический анализ гетероскедастичности

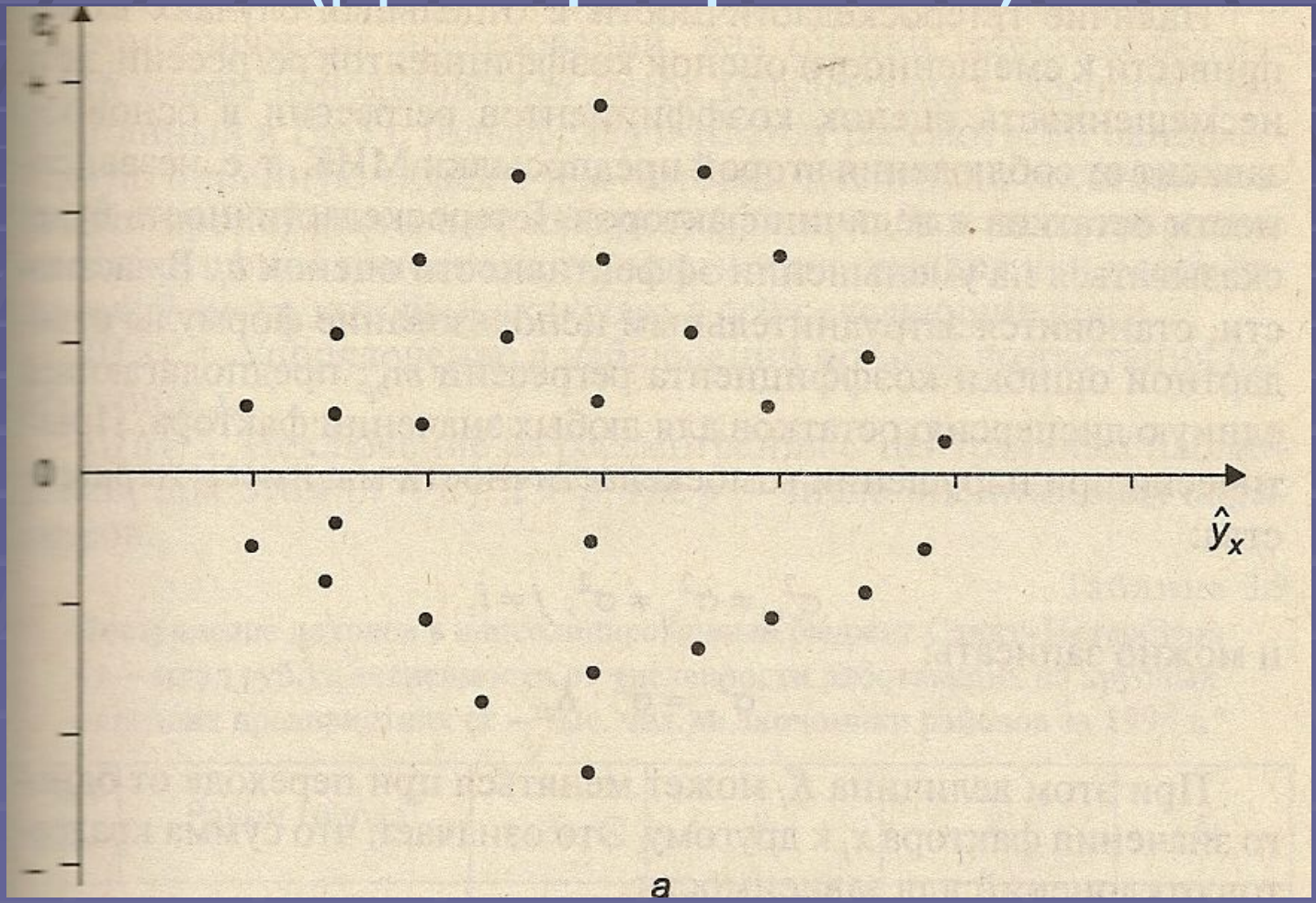


- а – дисперсия остатков растет при росте x
- б – дисперсия остатков при минимальном и максимальном значении x минимальна, при среднем значении x - максимальна

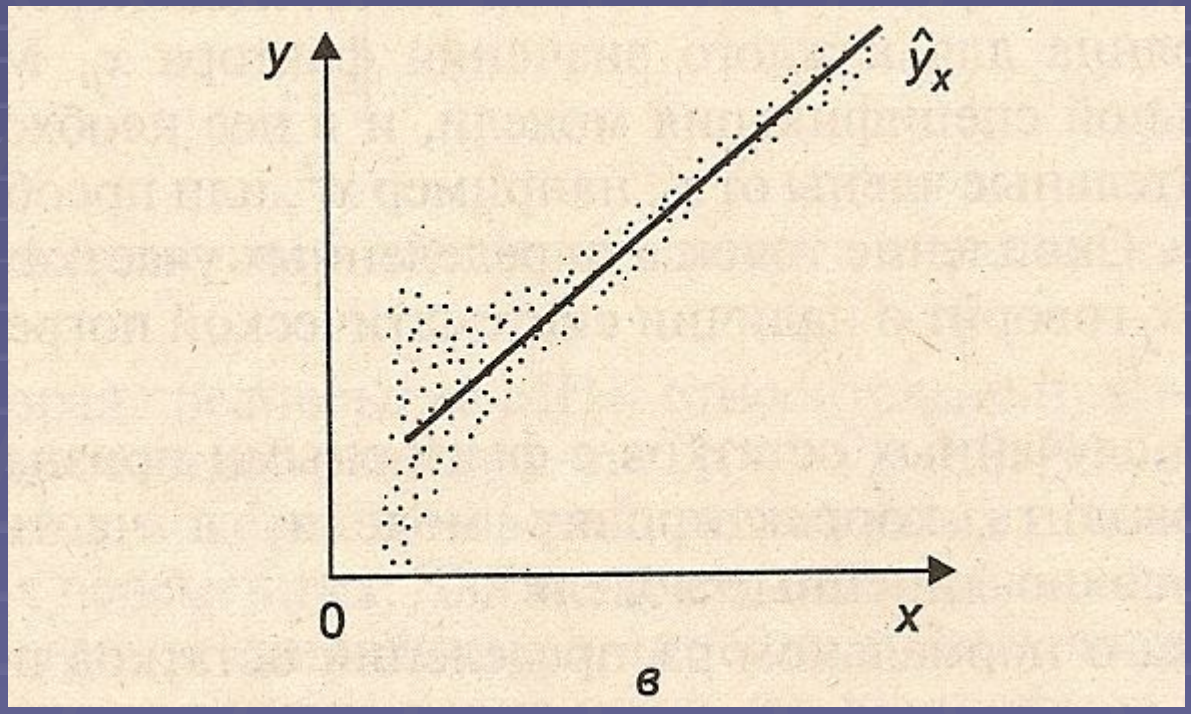
Графический анализ гетероскедастичности (для графика а)



Графический анализ гетероскедастичности (для графика б)

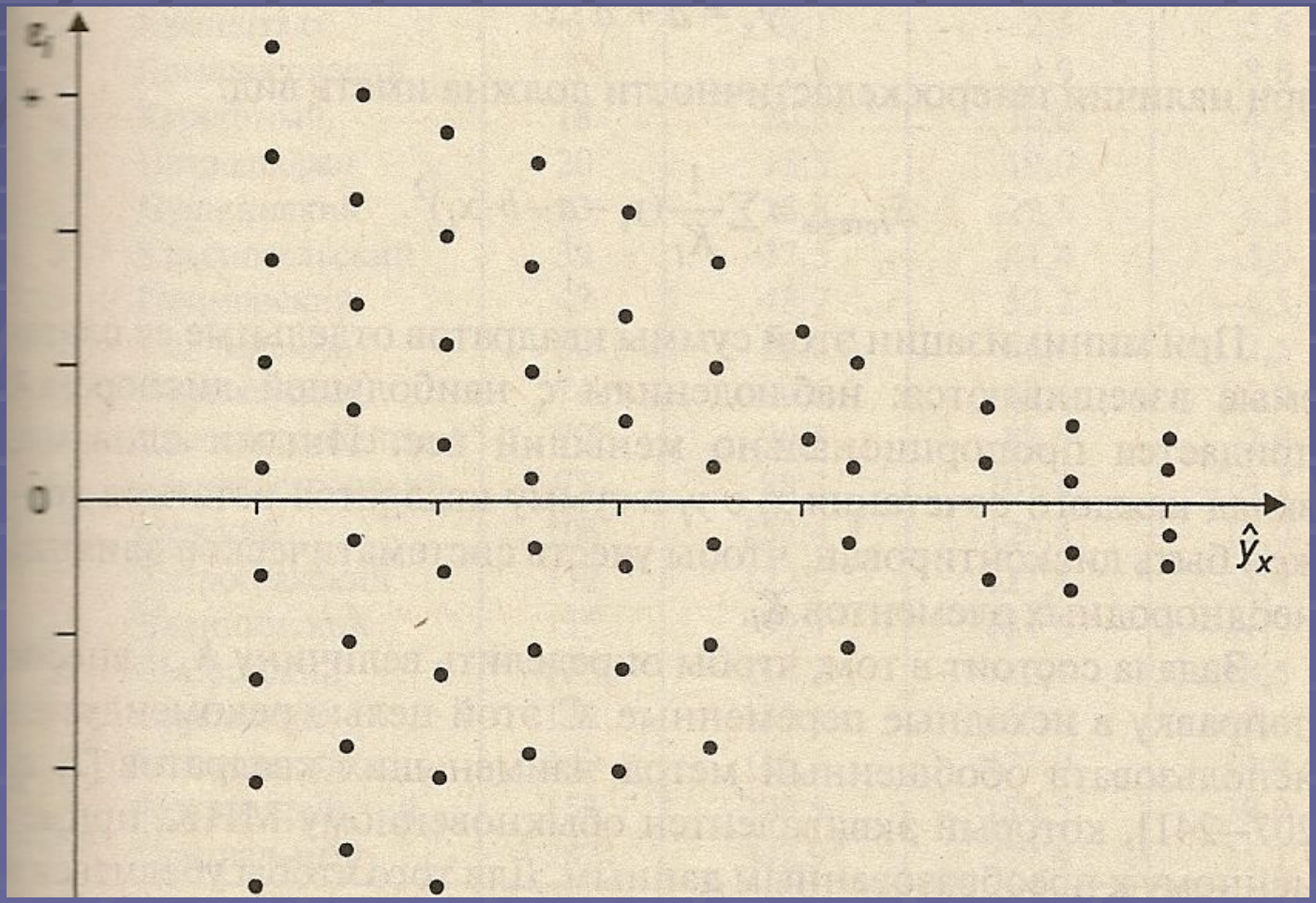


Графический анализ гетероскедастичности



- в – дисперсия остатков максимальна при минимальных значениях x

Графический анализ гетероскедастичности (для графика в)



Автокорреляция остатков

Для надежности результатов регрессии необходимо, чтобы автокорреляции остатков не было.

Её проверяют, например, на основе коэффициента автокорреляции r_a

$$r_a = \frac{\sum_{2}^{n} \varepsilon_i \varepsilon_{i-1}}{\sum_{1}^{n} \varepsilon_i^2}, \text{ если } r_a \text{ меньше табличного,}$$

то автокорреляции остатков нет

§ 6.2. КРИТЕРИЙ ДАРБИНА-УОТСОНА

Это наиболее известный способ обнаружения автокорреляции первого порядка. Пусть n — число наблюдений, m — число факторов модели, уровень значимости $\alpha = 0{,}05$. Для n, m, α по таблицам распределения Дарбина-Уотсона находим числа d_l и d_u.

Статистика Дарбина-Уотсона $DW = \dfrac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$.

Если $DW < d_l$, то это свидетельствует о положительной автокорреляции остатков. Если $DW > 4 - d_l$, то это свидетельствует об отрицательной автокорреляции остатков. При $d_u < DW < 4 - d_u$ гипотеза об отсутствии автокорреляции остатков принимается. Если $d_l < DW < d_u$ или $4 - d_u < DW < 4 - d_l$, то гипотеза об отсутствии автокорреляции остатков не может быть ни принята, ни отвергнута.

Ограничения при использовании критерия Дарбина-Уотсона:

1) $\beta_0 \neq 0$;

2) случайные отклонения определяются по авторегрессионной схеме первого порядка AR(1), то есть $\varepsilon_i = \rho\varepsilon_{i-1} + v_i$, где v_i — случайный член;

3) статистические данные должны иметь одинаковую периодичность (не должно быть пропусков в наблюдениях);

4) среди факторов не должно быть лаговых переменных (то есть переменных, влияние которых характеризуется определенным запаздыванием).

Пример использования DW

$\sum_{i=1}^{n} e_i^2 = 551{,}52.$

Номер	e_i	$e_i - e_{i-1}$	$(e_i - e_{i-1})^2$
1	8,3		
2	4,26	−4,04	16,32
3	−12,46	−16,72	279,56
4	−1,86	10,6	112,36
5	−7,38	−5,52	30,47
6	5,26	12,64	159,77
7	−9,66	−14,92	222,61
8	−2,26	7,4	54,76
9	8,34	10,6	112,36
10	7,46	−0,88	0,77
Сумма			988,98

Заполняем таблицу. Из каждого числа 2-го столбца вычитаем предыдущее число 2-го столбца и результат пишем в 3-м столбце. В 4-м столбце числа округляем до двух знаков после запятой. Статистика Дарбина-Уотсона:

$$DW = \frac{\sum_{i=2}^{2} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{988{,}98}{551{,}52} \approx 1{,}793.$$

По таблице распределения Дарбина-Уотсона находим $d_l = 0{,}697$ и $d_u = 1{,}641$. Тогда $4 - d_u = 4 - 1{,}641 = 2{,}359$.

Так как $d_u < DW < 4 - d_u$ $(1{,}641 < 1{,}793 < 2{,}359)$, то гипотеза об отсутствии автокорреляции остатков не отклоняется на уровне значимости 0,05. Это является одним из подтверждений высокого качества модели.

УСЛОВИЯ ИСПОЛЬЗОВАНИЯ УРАВНЕНИЙ РЕГРЕССИИ ДЛЯ ПРОГНОЗА

Если совокупность неоднородна по исследуемым признакам, то уравнение регрессии не имеет смысла

Должны быть неизменны условия формирования уровней признаков, которые лежат в основе определения оценок параметров модели регрессии.

Иначе необходимо собирать новый эмпирический материал, отражающий взаимосвязь признаков в новых условиях.

ПРИЗНАКИ ХОРОШЕЙ МОДЕЛИ

- Модель должна быть простой;
- Для любого набора статистических данных определяемые коэффициенты уравнения модели должны определяться однозначно;
- Стремятся строить модели с максимально возможным скорректированным коэффициентом детерминации R^2;
- Модель не может быть признана качественной, если она не соответствует известным теоретическим предпосылкам;
- Модель признается качественной, если полученные на её основе прогнозы подтверждаются реальностью.

ОШИБКИ СПЕЦИФИКАЦИИ

- это неправильный выбор функциональной формы модели или набора объясняющих переменных $x_1 \ldots x_p$

Основные их виды:

- Игнорирование значимой переменной (не включение её в модель);
- Добавление в модель незначимой переменной;
- Выбор неправильной функциональной формы

Любая качественная модель – подгонка спецификации модели под имеющиеся данные

- Из-за меняющихся условий протекания экономических процессов необходим постоянный пересмотр модели;
- При всех недостатках моделей принятие решений на их основе приводит к более точным результатам, чем принятие решений на основе интуиции и законов экономической теории

Thank You !