Регрессионный анализ

- линейная зависимость между переменными,

- метод наименьших квадратов и другие методы оценки коэффициентов,

- оценка эффективности коэффициентов уравнения регрессии

-простая и множественная регрессия,

-методы построения уравнений регрессии,

- методы анализа остатков.

ЗАРУБЕЖНЫЕ
СТАТИСТИЧЕСКИЕ
ИССЛЕДОВАНИЯ

ЗСИ

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ

•

Н. ДРЕЙПЕР
Г. СМИТ

APPLIED
REGRESSION ANALYSIS

N. R. DRAPER, H. SMITH

JOHN WILEY AND SONS, INC.
NEW YORK LONDON SYDNEY

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Н. ДРЕЙПЕР, Г. СМИТ

Перевод с английского, научное редактирование и предисловие **Ю. П. Адлера** и **В. Г. Горского**

«СТАТИСТИКА» МОСКВА 1973

СОДЕРЖАНИЕ

Химмельблау Д. Прикладное нелинейное программирование. М. – Мир., 1975. – 534 с.

И КОНТРОЛЮ ПРИРОДНОЙ СРЕДЫ

Государственный ордена Трудового Красного Знамени
гидрологический институт

на правах рукописи

ЛОБАНОВ Владимир Алексеевич

УДК 556.048

СТАТИСТИЧЕСКИЕ КРИТЕРИИ И РЕГРЕССИЯ ПРИ АНАЛИЗЕ МНОГОЛЕТНИХ КОЛЕБАНИЙ РЕЧНОГО СТОКА

05.14.09 – Гидравлика и инженерная гидрология

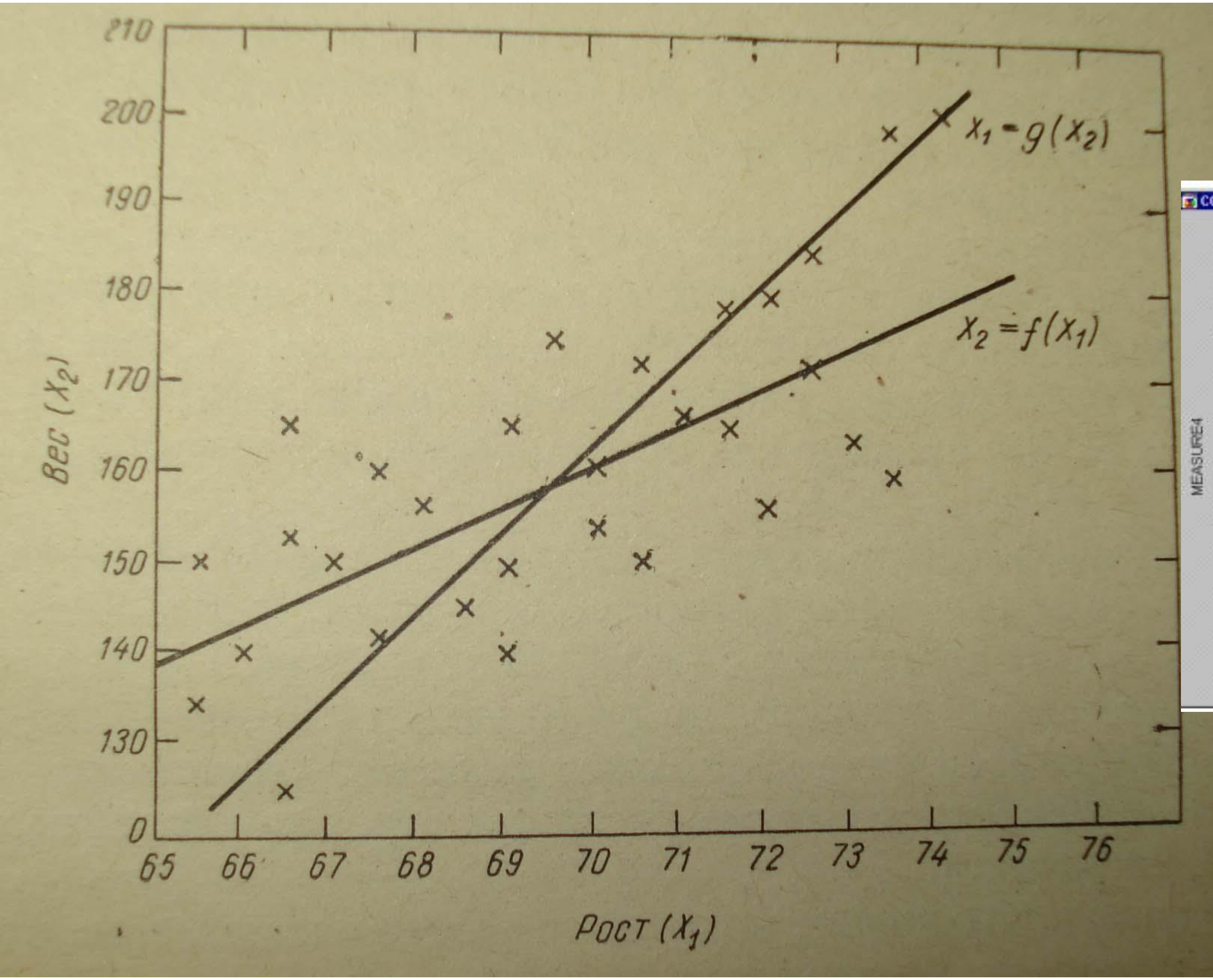
Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук
А.В.Рождественский

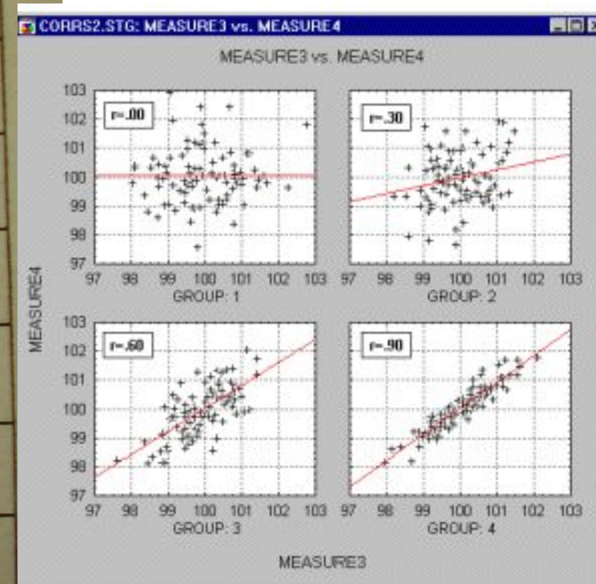
Ленинград – 1983

СОДЕРЖАНИЕ

Линейная зависимость между переменными

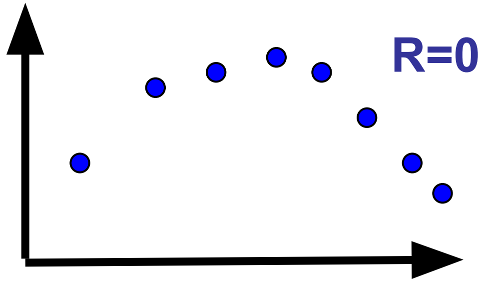


$$r = \frac{\sum (x_{1,i} - \bar{x}_1)\cdot(x_{2,i} - \bar{x}_2)}{\sqrt{\sum (x_{1,i} - \bar{x}_1)^2}\cdot\sqrt{\sum (x_{2,i} - \bar{x}_2)^2}}$$

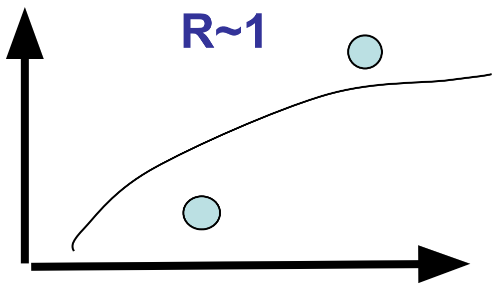


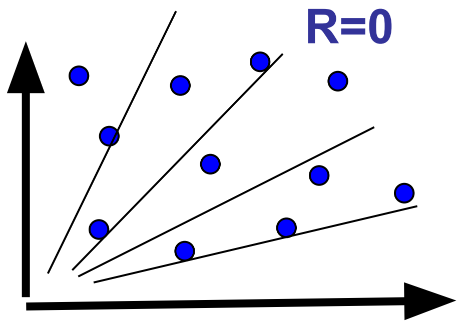
$$Y = b_1X + b_0 \pm \varepsilon$$

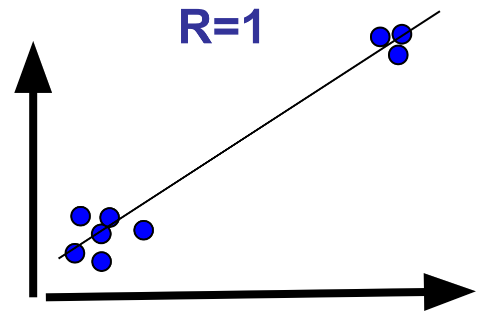
Линейная зависимость между переменными



Функциональные преобразования факторов





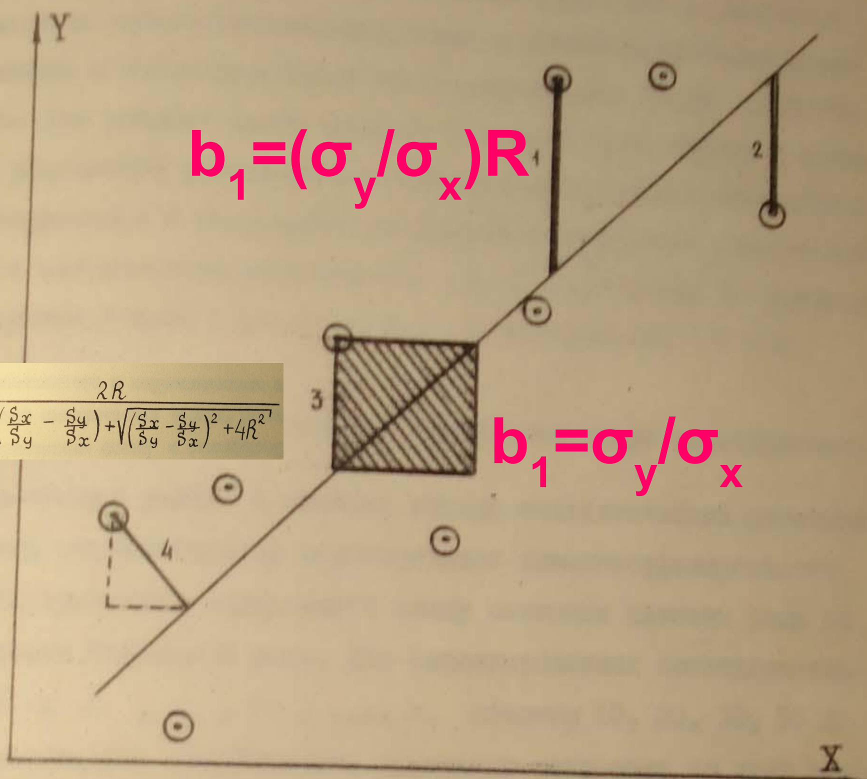


Статистическая незначимость

Совместный эффект

Ложная корреляция

Метод наименьших квадратов и другие методы оценки коэффициентов



$b_1=(\sigma_y/\sigma_x)R$

$$b_1'' = \frac{2R}{\left(\frac{S_x}{S_y} - \frac{S_y}{S_x}\right) + \sqrt{\left(\frac{S_x}{S_y} - \frac{S_y}{S_x}\right)^2 + 4R^2}}$$

$b_1=\sigma_y/\sigma_x$

Рис. 11. Геометрическая интерпретация методов расчета коэффициентов уравнения регрессии:

1 – МНК, $\min \Sigma \Delta Y_i^2$.

2 – метод абсолютных отклонений, $\min \Sigma |\Delta Y_i|$.

3 – метод одновначной аппроксимации, $\min \Sigma \Delta Y_i \Delta X_i$.

4 – метод ортогональной регрессии, $\min \Sigma \frac{\Delta Y_i \cdot \Delta X_i}{\sqrt{\Delta Y_i^2 + \Delta X_i^2}}$.

МНК

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \qquad (1.2.4)$$

Будем подбирать значения оценок b_0 и b_1 так, чтобы их подстановка вместо β_0 и β_1 в уравнение (1.2.4) давала наименьшее возможное (минимальное) значение S. (Заметим, что X_i, Y_i — это фиксированные числа, которые нам известны.) Мы можем определить b_0 и b_1 путем дифференцирования уравнения (1.2.4) сначала по β_0, затем по β_1 и приравнивания результатов к нулю. Тогда

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i),$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i), \qquad (1.2.5)$$

так что для оценок b_0 и b_1 имеем

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0,$$

$$\sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) = 0. \qquad (1.2.6)$$

где при приравнивании выражений (1.2.5) к нулю мы подставили (b_0, b_1) вместо (β_0, β_1). Из (1.2.6) имеем:

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0,$$

$$\sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0 \qquad (1.2.7)$$

или

$$b_0 n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i,$$

$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i. \qquad (1.2.8)$$

Эти уравнения называют *нормальными*.

Решение уравнений (1.2.8) относительно b_1 дает

$$b_1 = \frac{\Sigma X_i Y_i - [(\Sigma X_i)(\Sigma Y_i)]/n}{\Sigma X_i^2 - (\Sigma X_i)^2/n} = \frac{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma (X_i - \overline{X})^2}, \qquad (1.2.9)$$

$b_0=Ycp-b_1Xcp$

Оценка эффективности коэффициентов уравнения регрессии

Коэффициент B_1 уравнения

та.]
Стандартная ошибка b_1 есть корень квадратный из дисперсии, т. е.

$$\text{ст. ош. } (b_1) = \frac{\sigma}{\{\Sigma (X_i - \overline{X})^2\}^{1/2}}.$$

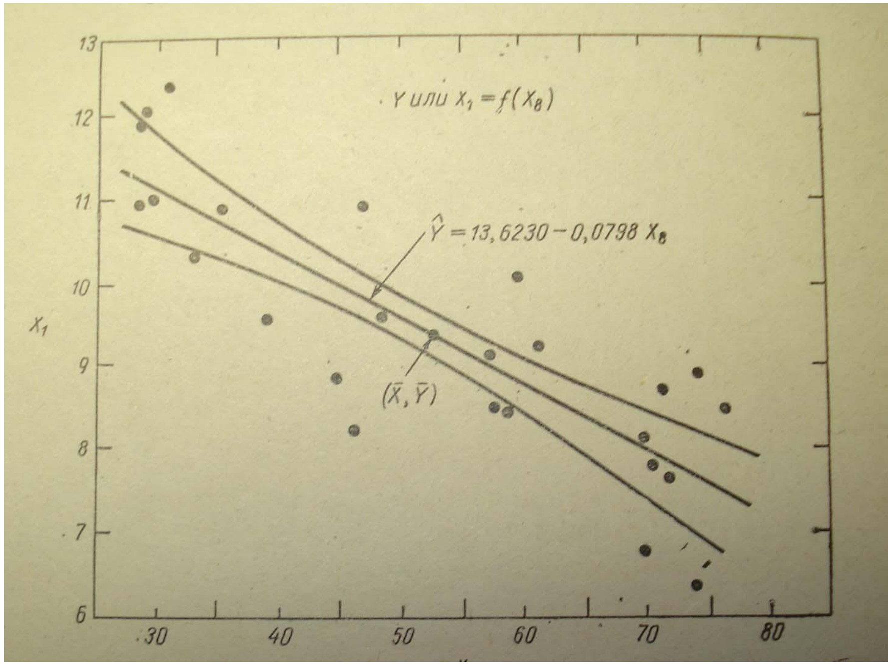
Если σ неизвестна и мы применяем вместо нее оценку s, предполагая, что модель корректна, то *оценка* стандартной ошибки b_1 есть

$$\text{оц. ст. ош. } (b_1) = \frac{s}{\{\Sigma (X_i - \overline{X})^2\}^{1/2}}. \qquad (1.4.2)$$

Если мы предполагаем, что разброс наблюдений относительно ли-
ии нормален, т. е., что ошибки ε_i все принадлежат некоторому нор-
альному распределению, $N(0, \sigma^2)$, то можно показать, что 100 (1 —
– α)%-ные доверительные интервалы для β_1 получатся, если вычис-
ить

$$b_1 \pm \frac{t\left(n-2, 1-\frac{1}{2}\alpha\right) s}{\{\Sigma (X_i - \overline{X})^2\}^{1/2}}, \qquad (1.4.3)$$

Доверительные интервалы уравнения регрессии



Коэффициент B_0 уравнения

$$\text{кв. ош. } (b_0) = \left\{\frac{\Sigma X_i^2}{n\Sigma (X_i - \overline{X})^2}\right\}^{1/2} \sigma.$$

Отсюда имеем 100 (1 — α)%-ные доверительные пределы для β_0

$$b_0 \pm t\left(n-2,\ 1-\frac{1}{2}\alpha\right)\left\{\frac{\Sigma X_i^2}{n\Sigma (X_i - \overline{X})^2}\right\}^{1/2} s.$$

Расчетное значение Yp

$$\text{оц. кв. ош. } (\hat{Y}_k) = s\left\{\frac{1}{n} + \frac{(X_k - \overline{X})^2}{\Sigma (X_i - \overline{X})^2}\right\}^{1/2}.$$

Простая и множественная регрессия

$Y = b_1X + b_0 \pm \varepsilon$

$Y = b_1X_1 + b_2X_2 + \ldots + b_0 \pm \varepsilon$ аддитивная

$Y = b_1{*}X_1{*}X_2 + \ldots + b_0 \pm \varepsilon$ мультипликативная

$Y = b_1X + b_2{*}X_1{*}X_2 + \ldots + b_0 \pm \varepsilon$ смешанная с совместными эффектами

$Y = b_1X^2 + b_2{*}X_1{*}X_2 + b_3\log X_3 + \ldots + b_0 \pm \varepsilon$

смешанная с совместными эффектами и функциональными преобразованиями факторов

Методы построения уравнений регрессии

1. Метод исключения

Алгоритм:

-рассчитывается регрессионное уравнение, включающее все переменные (факторы);

-рассчитывается величина частного *F*-критерия для каждой из рассматриваемых
-переменных, как будто бы она была последней переменной, введенной в регрессионное уравнение;

-наименьшая величина частного *F*-критерия, обозначаемая через *FL*, сравнивается с заранее выбранным уровнем значимости (*F*0);

-если *FL*<*F*0, то переменная *XL* , связанная с *FL*, исключается из рассмотрения и производится перерасчет уравнения регрессии с учетом оставшихся переменных, затем снова рассчитывается величина частного *F*-критерия для каждой из оставшихся переменных и процедура повторяется;

- если *FL*>*F*0, то остается то уравнение, которое построено.

$$F_{част_{x_i}} = \frac{R^2_{yx_1 \cdots x_i \cdots x_p} - R^2_{yx_1 \cdots x_{i-1} x_{i+1} \cdots x_p}}{1 - R^2_{yx_1 \cdots x_i \cdots x_p}} \cdot \frac{n-m-1}{1}$$

Частный *F*-критерий - отношение среднего квадрата новой переменной Xi с одной степенью свободы к дисперсии модели.

Методы построения уравнений регрессии

2. Шаговая процедура

Включаются все переменные по очереди до тех пор, пока регрессионное уравнение не станет удовлетворительным.

Алгоритм:

-прежде всего выбирается переменная X1, имеющая наиболее высокий коэффициент парной корреляции с Y и определяется расчетное значение Yp по однофакторному уравнению с X1;

- определяется частный коэффициент корреляции между остатками E=Y-Yp и остальными переменными за исключением X1;

-выбирается величина X2, которая имеет наибольший частный коэффициент корреляции и находится второе регрессионное уравнение Y=f(X1, X2);

-далее включенная переменная X2 исследуется на эффективность включения по частному *F*-критерию таким же образом, как и в методе исключения;

- также, если $FL>F0$, то переменная остается в уравнении, если $FL<F0$, то исключается и процедура повторяется для следующей переменной с наибольшим частным коэффициентом корреляции.

Оценка эффективности уравнения

Фундаментальное уравнение регрессионного анализа

Уравнение (1.3.2) можно записать иначе[12]:

$$\Sigma (Y_i - \bar{Y})^2 = \Sigma (Y_i - \hat{Y}_i)^2 + \Sigma (\hat{Y}_i - \bar{Y})^2. \quad (1.3.3)$$

Теперь $Y_i - \bar{Y}$ — это отклонение i-го наблюдения от общего среднего, следовательно, левая часть уравнения (1.3.3) — это сумма квадратов отклонений наблюдений от среднего; сокращенно — *SS относительного среднего*, а также *скорректированная сумма квадратов Y-ов*. Так как $Y_i - \hat{Y}_i$ есть отклонение i-го наблюдения от его предсказанного или вычисленного значения (i-й *остаток* — см. гл. 3), а $\hat{Y}_i - \bar{Y}$ есть отклонение предсказанного значения i-го наблюдения от среднего, то мы можем выразить уравнение (1.3.3) следующим образом[12]:

$$\begin{pmatrix}\text{Сумма квадратов}\\ \text{относительно среднего}\end{pmatrix} = \begin{pmatrix}\text{Сумма квадратов}\\ \text{относительно регрес-}\\ \text{сии}\end{pmatrix} + \begin{pmatrix}\text{Сумма квадратов, обу-}\\ \text{словленная регрессией}\end{pmatrix}^{12А}$$

[12] Уравнение (1.3.3) играет фундаментальную роль в дисперсионном анализе. Можно даже сказать, что в нем в зародыше содержится весь дисперсионный анализ. — *Прим. перев.*

Таблица дисперсионного анализа

Источник	Сумма квадратов	Степень свободы	Средний квадрат
Регрессия	$b_1 \left\{ \Sigma X_i Y_i - \dfrac{(\Sigma X_i)(\Sigma Y_i)}{n} \right\}$	1	MS_R
Относительно регрессии (остаток)	По разности	$n-2$	$s^2 = \dfrac{(SS)}{(n-2)}$
Относительно среднего (общий, скорректированный на среднее)	$\Sigma Y_i^2 - \dfrac{(\Sigma Y_i)^2}{n}$	$n-1$	

Методы анализа остатков

Остатки любой эмпирической зависимости определяются как разности между фактическими (наблюденными) и расчетными значениями:

εi=Yi-Ÿi, или *Δi =(Yi-Ÿi)/ Yi ,*

где: Yi - фактическое (наблюденное) значение, *Ÿi* - рассчитанное по зависимости, *εi* – остаток или погрешность полученной зависимости.

Наиболее распространенной обобщенной характеристикой остатков является их среднее квадратическое отклонение (σ*i*):

$$\sigma_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}{n-1}}$$

или:

$$\sigma_\varepsilon = \sigma_Y \sqrt{1 - R^2}$$

В качестве обобщенной меры может служить также величина:

$$\Delta' = (1 - R^2) * 100\%,$$

которая характеризует долю исходного рассеивания (в %), не объясненного с помощью построенной зависимости.

Методы анализа остатков

Всесторонний анализ остатков включает в себя оценивание:
- резко отклоняющихся экстремальных значений;
- смещенности остатков;
- случайности остатков на основе хронологического графика;
- случайности остатков в зависимости от каждого фактора, входящего в уравнение;
- случайности остатков в зависимости от расчетных значений.

Смещенность

Наличие смещенности остатков определяется тем, что среднее их значение не равно нулю или статистически значимо отличается от нуля.

Как правило, наличие смещенности может иметь место для уравнений балансового вида (уравнение водного, руслового и других видов баланса), где невязки уравнений характеризуют как неучтенные факторы, так и все систематические погрешности составляющих.

Смещенность остатков необходимо исключать или путем корректировки свободного члена уравнения или тех коэффициентов и факторов, которые ее обусловили.

Методы анализа остатков

Случайность остатков во времени

Два пути:

-применение известных статистических критериев оценки случайности и стационарности (критерии Стьюдента, Фишера и другие);

- применение графического анализа остатков в зависимости от времени.

При неслучайном характере возможны следующие основные варианты:

- полоса разброса остатков сужается или расширяется, что связано с непостоянством дисперсии остатков во времени;

-полоса остатков имеет одинаковую ширину, но изменяется (линейно или нелинейно) в зависимости от времени, что свидетельствует о нестационарности средних значений остатков.

Случайность остатков от факторов и расчетного значения

Возможны следующие ситуации:

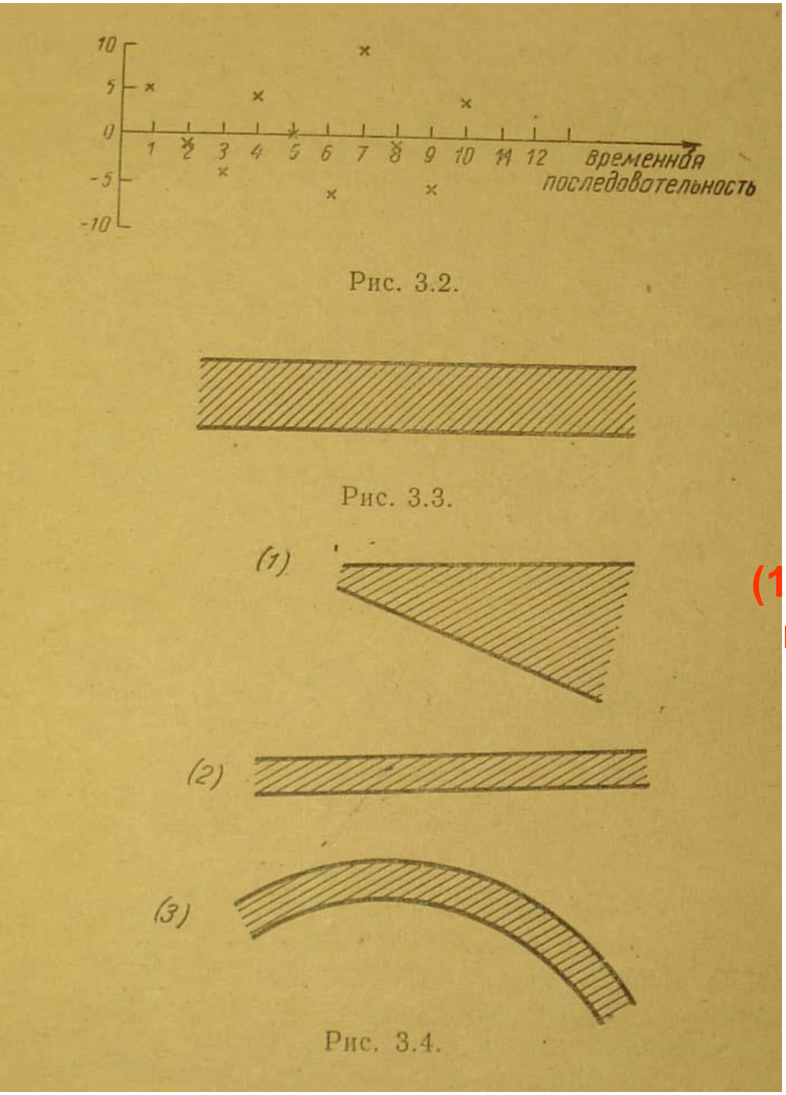
- зависимость отсутствует и полоса остатков горизонтальна и симметрична относительно нулевого значения, что свидетельствует о случайности погрешностей;
- зависимость представлена сужающейся или расширяющейся полосой остатков от фактора, что свидетельствует о неоднородности дисперсии остатков, которую надо учитывать взвешенным МНК или предварительным преобразованием Y_i;
- линейная зависимость остатков от фактора свидетельствует о том, что линейный эффект данного фактора в уравнении исключен неверно;
- нелинейная зависимость остатков от фактора свидетельствует о том, что в уравнение необходимо ввести нелинейные члены от X_i или произвести преобразование Y_i.

Проверка построенного эмпирического уравнения на независимом от расчета материале наблюдений.

Анализ остатков в случае независимой проверки осуществляется теми же способами: на резко отклоняющиеся экстремумы, в зависимости от времени, факторов и расчетного значения.

Необходимо отметить, что должен иметь место оптимум между количеством информации, используемой для построения зависимости и для ее независимой проверки.

Методы анализа остатков



(1) Дисперсия не постоянна, а расчетт со временем – следует использовать взвешенный МНК

(2) В модель следует следует включить фактор времени

(3) В модель должны быть включены линейный и квадратичный члены от времени.

ПРИМЕР Оценка эффективности эмпирической зависимости

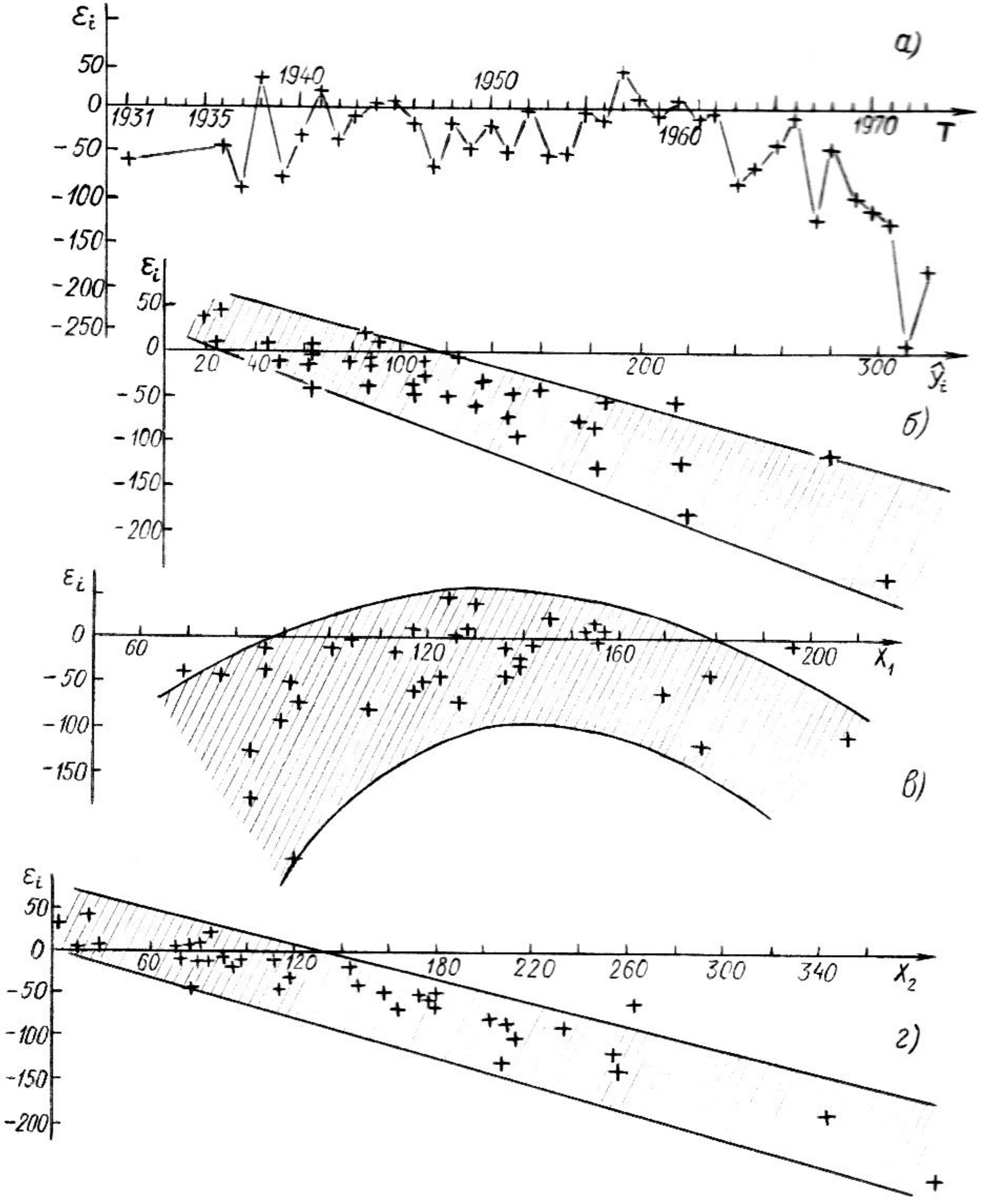
Слой поверхностного стока весеннего половодья (Y) на р.Оке – с.Половское

$$Y = 0.80X_1 + 0.86X_2 – 104,$$

где: X1 = S + Sл + Xос ,
S - максимальные запасы воды в снеге (мм), Sл – запас воды в ледяной корке (мм), Xос – осадки за период половодья (мм);

X2 = (L * e)/50, L – глубина промерзания почвы (см), e – величина осеннего увлажнения почвы (см).

Анализ остатков эмпирической зависимости для расчета слоев половодья



зависимость остатков от времени (ε=f(t))

зависимость остатков от расчетных слоев стока (ε=f(Y')

зависимость остатков от первого фактора (ε=f(X1)

зависимость остатков от второго фактора (ε=f(X2)

Выводы:

-с 1967 г. остатки зависят от времени и имеет место существенное систематическое завышение слоя стока половодья, вычисленного по эмпирической зависимости (рис.1а);

-наклонная полоса рассеяния на рис.1б показывает, что отклонения от полученной эмпирической зависимости носят систематический характер: отрицательные остатки соответствуют большим по величине значениям расчетных слоев стока, положительные – малым, что свидетельствует о неточном определении свободного члена в уравнении;

-изгиб полосы рассеяния на рис.1в показывает, что в уравнении необходимо учесть нелинейность зависимости Y от X_1;

- из рис.1г следует, что коэффициент перед X_2 также определен неверно.

Новая эмпирическая зависимость:

$$\mathbf{Y = 0.76X'_1 + 0.14*10\text{-}5X^2_2 + 13.8}$$

где: $X'_1 = X_{10} * \gamma_п$ и

X_{10} – средний максимальный снегозапас в бассейне (мм), осредненный по 10 метеостанциям, для которых коэффициент корреляции снегозапасов со стоком половодья >0.5;

$\gamma_п$ = Кп * βп и

Кп – модульный коэффициент приведенных запасов влаги в почве,

βп = sin (α'+100), tg α'= $К_Н$ и

$К_Н$ - модульный коэффициент промерзания;

X_2 – слой стока за март (мм).

Коэффициент корреляции полученного уравнения равен 0.89