Мультиколлинеарность: понятие, признаки, методы устранения

Мультиколлинеарность

Мультиколлинеарность – совместная, или множественная, взаимозависимость объясняющих переменных. Выделяют:

1. **Полную мультиколлинеарность** - если существует функциональная линейная зависимость между объясняющими переменными, определяется нарушением одного из требований КЛММР, а именно, требования к рангу матрицы X .

2. **Реальная (или частичная) мультиколлинеарность** возникает в случаях существования достаточно тесных линейных статистических связей между объясняющими переменными.

Полная мультиколлинеарность и ее последствия

Когда между объясняющими переменными существует функциональная зависимость (**полная мультиколлинеарность**), то матрица объясняющих переменных (матрица X) вырождена, так как ранг матрицы X меньше *к+1*, что в свою очередь, приводит к вырожденности $X^T X$,а это значит не существует обратная к этой матрице, и следовательно невозможно оценить коэффициенты методом наименьших квадратов.

Полную мультиколлинеарность несложно избежать уже на предварительной стадии анализа и отбора множества объясняющих переменных, например в матрице $X^T X$, определяя ее ранг (например, методом элементарных преобразований) выявить, какие строки линейно зависят от других строк. Выявив эти строки, из модели линейной регрессии исключаются соответствующие этим строкам объясняющие переменные. Таким образом, строится регрессионную модель меньшей размерности, а по оставшимся объясняющим переменным матрица $X^T X$ будет невырожденна.

Реальная (частичная) мультиколлинеарность

При сильной статистической связи объясняющих переменных (**реальной мультиколлинеарности**) матрица X^TX становится плохо (слабой) обусловленной – ее определитель близок к нулю. При этом элементы матрицы $(X^TX)^{-1}$ вычисляются с большой погрешностью, следовательно, и оценки, полученные МНК, тоже определены с погрешностью. Это, как правило, влечет за собой и увеличение дисперсии оценок коэффициентов в регрессионной модели, следовательно, и процедуры проверки существенности параметров, которые будут давать смещенные результаты. Такое смещение означает, что в ряде случаев наш вывод, согласно которому некоторые оценки незначительно отличаются от нуля, будет ложным.

Внешние признаки реальной мультиколлинеарности

- неоправданно большие с экономической точки зрения значения оценок коэффициентов уравнения регрессии;
- небольшие изменения исходных статистических данных приводит к существенному изменению оценок коэффициентов модели, вплоть до изменения их знаков;
- неправильные с экономической точки зрения знаки отдельных коэффициентов регрессии;
- среди коэффициентов уравнения регрессии много (может быть все) незначимы, а модель значима;
- стандартные отклонения велики на столько, что сравнимы или даже превосходят сами коэффициенты;
- доверительные интервалы для коэффициентов уравнения регрессии содержат внутри себя точку ноль.

Формальные признаки мультиколлинеарности

- среди значимых коэффициентов парной или частной корреляции объясняющих переменных есть такие, которые по абсолютной величине достаточно велики (превышают 0,75-0,8);
- достаточно высокие значение множественного коэффициента корреляции (детерминации) одной из объясняющих переменных на другие ;
- **необходимым условием** плохой обусловленности является малость определителя матрицы $X^T X$. Если значение оказывается близким к нулю, то свидетельствует о наличии мультиколлинеарности.
- **достаточным условием** плохой обусловленности (мультиколлинеарности) является большое значение числа обусловленности.

$$M = \frac{\max|\lambda_i|}{\min|\lambda_i|} \quad i = \overline{1,n}, \text{ где } \lambda_i \text{ - собственное число матрицы } X^T X$$

Анализ внешних признаков мультиколлинеарности

Оценка линейной функции множественной регрессии, описывающей зависимость y – урожайности зерновых культур (ц/га) имеет вид:

N=20	Regression Summary for Dependent Variable: Y (Spreadsheet1) R= ,71927819 R?= ,51736112 Adjusted R?= ,34499008 F(5,14)=3,0014 p<,04784 Std.Error of estimate: 1,5989					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(14)	p-level
Intercept			3,52837	5,40597	0,652679	0,524536
X1	-0,011492	0,997196	-0,01068	0,92679	-0,011525	0,990967
X2	0,358696	0,496636	15,48714	21,44289	0,722251	0,482031
X3	0,156603	1,133826	0,11423	0,82705	0,138119	0,892113
X4	0,728655	0,251310	4,47482	1,54334	2,899432	0,011655
X5	-0,288196	0,303399	-2,92747	3,08191	-0,949890	0,358279

$$\underset{(5,41)}{\hat{y} = 3,52} \underset{(0.92)}{- 0,01} \cdot x_1 \underset{(21,44)}{+ 15,49} \cdot x_2 \underset{(0,83)}{+ 0,11} \cdot x_3 \underset{(1.54)}{+ 4,48} \cdot x_4 \underset{(3,08)}{- 2,93} \cdot x_5$$

$$F_{набл} = 3,01 \qquad F_{kp}(0,05;5;14) = 2.982 \qquad t_{kp}(0.05;14) = 2.145$$

1. Значения оценок коэффициентов уравнения регрессии соответствуют значениям исходным статистическим данным.

2. Небольшие изменения исходных статистических данных (на 0,5) привели к изменением оценок коэффициентов (в 1,3 р).

	1 Y	2 X1	3 X2	4 X3	5 X4	6 X5
1	10,2	2,09	0,76	2,55	0,82	0,64
2	8,9	0,84	0,78	0,96	1,09	1,16
3	9,5	3,03	0,81	2,96	0,8	0,81
4	10,4	5,13	0,9	6,94	0,93	1,09
5	10,1	2,66	0,76	2,66	0,89	0,66
6	9,1	2,66	0,8	3,19	0,82	0,67
7	13	1,18	0,79	1,23	0,92	0,73
8	8,1	0,85	0,76	0,92	0,71	0,58
9	7,4	1,02	0,74	0,99	0,7	0,58
10	14	3,92	0,81	3,52	1,87	1,23
11	10,2	2,28	0,8	3,69	1,23	0,67
12	11,2	2,9	0,82	3,8	0,75	0,64
13	12,6	9,86	0,9	12,01	0,89	0,88
14	10,2	2,28	0,78	2,76	1,32	0,67
15	7,5	1,09	0,79	1,1	0,63	0,85
16	7,7	0,78	0,76	0,8	0,59	0,65
17	8,7	2,14	0,79	1,94	0,7	0,58
18	8,9	0,59	0,72	0,55	0,93	0,7
19	13,6	0,58	0,75	0,8	1,23	0,7
20	9,2	1,86	0,76	0,67	1,49	0,92

Regression Summary for Dependent Variable: Y (Spreadsheet1)
R= ,72290267 R?= ,52258827 Adjusted R?= ,35208409
F(5,14)=3,0650 p<,04485 Std.Error of estimate: 1,5902

N=20	Beta	Std.Err. of Beta	B	Std.Err. of B	t(14)	p-level
Intercept			-2,44770	15,28816	-0,160104	0,875086
X1	-0,271828	0,977439	-0,25266	0,90852	-0,278102	0,784999
X2	0,287999	0,494213	12,43473	21,33827	0,582743	0,569334
X3	0,471819	1,111022	0,34558	0,81375	0,424671	0,677530
X4	0,729121	0,250250	4,47768	1,53683	2,913571	0,011334
X5	-0,254533	0,303342	-2,58553	3,08133	-0,839097	0,415513

3. Наличие незначимых коэффициентов при переменных X1, X2, X3, X5, при этом модель в целом значима;

4. Стандартные ошибки оценок коэффициентов при переменных X1, X2, X3, X5 превосходят значения самих коэффициентов;

5. Содержательно неинтерпретируемые коэффициенты при переменных X1, X2, X3, X5 .

N=20	Regression Summary for Dependent Variable: Y (Spreadsheet1) R= ,71927819 R?= ,51736112 Adjusted R?= ,34499008 F(5,14)=3,0014 p<,04784 Std.Error of estimate: 1,5989					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(14)	p-level
Intercept			3,52837	5,40597	0,652679	0,524536
X1	-0,011492	0,997196	-0,01068	0,92679	-0,011525	0,990967
X2	0,358696	0,496636	15,48714	21,44289	0,722251	0,482031
X3	0,156603	1,133826	0,11423	0,82705	0,138119	0,892113
X4	0,728655	0,251310	4,47482	1,54334	2,899432	0,011655
X5	-0,288196	0,303399	-2,92747	3,08191	-0,949890	0,358279

$$\hat{y} = \underset{(5,41)}{3,52} - \underset{(0.92)}{0,01}\cdot x_1 + \underset{(21,44)}{15,49}\cdot x_2 + \underset{(0,83)}{0,11}\cdot x_3 + \underset{(1.54)}{4,48}\cdot x_4 - \underset{(3,08)}{2,93}\cdot x_5$$

$$t_{kp}(0.05;14) = 2.145$$

$x^{(1)}$-Число тракторов (приведённой мощности) на 100 га;

$x^{(2)}$ -Число зерноуборочных комбайнов на 100 га;

$x^{(3)}$ -Число орудий поверхностной обработки почвы на 100 га;

$x^{(4)}$ - Количество удобрений, расходуемых на гектар (т/га);

$x^{(5)}$ -Количество средств химической защиты растений, расходуемых на гектар (ц/га);

Анализ формальных признаков мультиколлинеарности

1. Наличие значимых парных коэффициентов корреляции (превосходящих 0,75) между объясняющими переменными X1 и X2; X1 и X3, X2 и X3

Variable	**X1**	X2	X3	X4	X5
X1	1,0000	,8543	,9778	,1104	,3410
	p= ---	p=,000	p=,000	p=,643	p=,141
X2	,8543	1,0000	,8818	,0269	,4596
	p=,000	p= ---	0	p=,911	p=,041
X3	,9778	,8818	1,0000	,0303	,2784
	0	0	p= ---	p=,899	p=,235
X4	,1104	,0269	,0303	1,0000	,5706
	p=,643	p=,911	p=,899	p= ---	p=,009
X5	,3410	,4596	,2784	,5706	1,0000
	p=,141	p=,041	p=,235	p=,009	p= ---

Рисунок 1 - Матрица парных коэффициентов корреляции между факторными признаками

$x^{(1)}$-Число тракторов (приведённой мощности) на 100 га;

$x^{(2)}$ -Число зерноуборочных комбайнов на 100 га;

$x^{(3)}$ -Число орудий поверхностной обработки почвы на 100 га;

$x^{(4)}$ - Количество удобрений, расходуемых на гектар (т/га);

$x^{(5)}$ -Количество средств химической защиты растений, расходуемых на гектар (ц/га);

2. Наличие высоких коэффициентов детерминации $R^2_{x_j / x_1 .. x_{j-1}, x_{j+1}, .., x_k}$

$$R^2_{x_1 / x_2, x_3, x_4, x_5} = 0{,}97$$

$$R^2_{x_2 / x_1, x_3, x_4, x_5} = 0{,}86$$

$$R^2_{x_3 / x_1, x_2, x_4, x_5} = 0{,}98$$

$$R^2_{x_4 / x_1, x_2, x_3, x_5} = 0{,}45$$

$$R^2_{x_5 / x_1, x_2, x_3, x_4} = 0{,}62$$

$x^{(1)}$-Число тракторов (приведённой мощности) на 100 га;

$x^{(2)}$ -Число зерноуборочных комбайнов на 100 га;

$x^{(3)}$ -Число орудий поверхностной обработки почвы на 100 га;

$x^{(4)}$ - Количество удобрений, расходуемых на гектар (т/га);

$x^{(5)}$ -Количество средств химической защиты растений, расходуемых на гектар (ц/га);

3. Малость определителя матрицы $X^T X$

$$\left| x^T x \right| = 41.502$$

4. Большое значение числа обусловленности матрицы $X^T X$

$$cond\left(X^T X\right) = \frac{\max|\lambda_j|}{\min|\lambda_j|} = \frac{402{,}318}{5{,}19 \times 10^{-3}} = 7{,}752 \times 10^4$$

Методы устранения мультиколлинеарности

1. . Метод пошаговой регрессии :

- с включением переменных
- с исключением переменных

2. Метод «ридж-регрессии

3. Метод главных компонент

Метод пошаговой регрессии

1. Метод пошаговой регрессии с включением переменных

Решается **задача**: для заданного значения l $\left(l=\overline{1,k-1}\right)$ путем перебора возможных комбинаций из l объясняющих переменных, отобранных из исходного набора $x_1, x_2, \ldots x_{\hat{e}}$, определить такие переменные $x^{\left(i_1^0(l)\right)}, x^{\left(i_2^0(l)\right)}, \ldots, x^{\left(i_l^0(l)\right)}$, для которых коэффициент детерминации с результирующим показателем y был бы максимальным.

$$\hat{R}^2_{y/x^{i_1^0(l)},\ldots,x^{i_l^o(l)}} = \max_{1\le i_1,i_2,\ldots i_l\le k} \hat{R}^2_{y/x^{i_1},\ldots,x^{i_l}}$$

На **первом шаге** процедуры ($l=1$) определяется одна (первая) объясняющая переменная $x^{(i_1(1))}$, которую можно назвать наиболее информативной, при условии, что в регрессионную модель Y по X включена только одна из набора объясняющих переменных.

На **втором шаге** реализация критерия максимальности коэффициента детерминации определит уже наиболее информативную пару переменных $x^{(i_1(1))}, x^{(i_2(2))}$, при этом одна из них определена на предыдущем (первом) шаге. Эта пара будет иметь наиболее тесную статистическую связь с результирующим показателем Y.

На **третьем шаге** ($l=3$) будет отобрана наиболее информативная тройка объясняющих переменных и т.д.

На каждом шаге определяются несмещенная оценка коэффициента детерминации

$$\overline{R}^{*2}(l) \cong 1 - \left(1 - \overline{R}^{2}(l)\right)\frac{n-1}{n-l-1}.$$

и оценка нижней доверительной границы $\overline{R}^{2}_{\min}(l)$ для $\overline{R}^{2}(l)$

$$\overline{R}^{2}_{\min}(l) = \overline{R}^{*2}(l) - 2\sqrt{\frac{2l(n-l-1)}{(n-1)(n^2-1)}}\left(1 - \overline{R}^{2}(l)\right).$$

Правило отбора объясняющих переменных: **предполагается выбирать в качестве оптимального числа l_0 объясняющих переменных регрессионной модели значение l, при котором величина $R^2_{\min}(l)$ достигает своего максимума.**

Метод пошаговой регрессии с исключением переменных

На **первом шаге** строится уравнение регрессии на все k факторных признаков и, если среди его коэффициентов есть незначимые, то **на втором шаге** строятся уравнения регрессии на $k-1$ признаков, среди которых выбирается то, которому соответствует наибольший выборочный коэффициент детерминации. Если и в этой модели есть незначимые коэффициенты, то процедура повторяется для $k-2$ переменных и т. д.

Пример реализации метода пошаговой регрессии

Шаг 1. Выбираем первую переменную, включаемую в модель:

$$R^2_{y/x1} = 0.185 \quad R^2_{y/x3} = 0.163 \quad R^2_{y/x5} = 0.11$$

$$R^2_{y/x2} = 0.139 \quad R^2_{y/x4} = 0.333$$

$$\max_{i=\overline{1,5}} \overline{R}^2_{y/x_i} = \overline{R}^2_{y/x_4} = 0{,}333$$

Поправленный на несмещённость коэффициент детерминации:

$$\overline{R}^{*2}_{y/x4} = 0.296$$

Нижняя граница доверительного интервала

$$R^2_{\min}(1) = 0.204$$

Шаг 2. Выбираем вторую переменную, включаемую в модель:

$$R^2_{y/x4,x1} = 0.469$$

$$R^2_{y/x4,x2} = 0.462$$

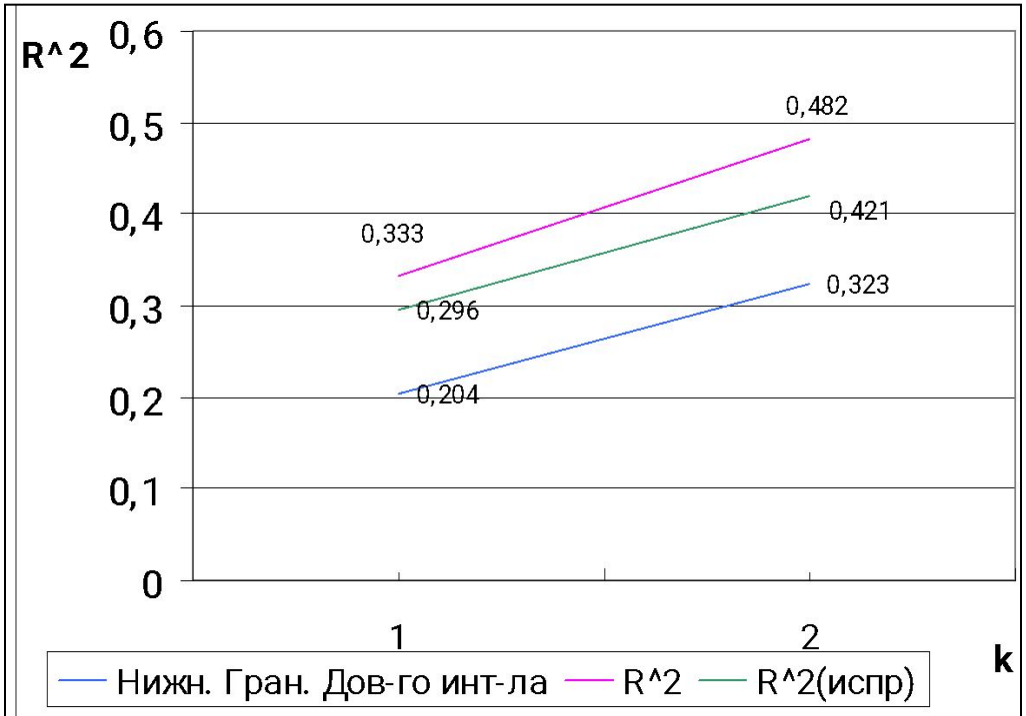
$$R^2_{y/x4,x3} = 0.482$$

$$R^2_{y/x4,x5} = 0.333$$

$$\max_{i=\overline{1,2,3,5}} \bar{R}^2_{y/x_4,x_i} = \bar{R}^2_{y/x_4,x_3} = 0{,}482$$

$$\bar{R}^{*2}_{y/x4,x3} = \bar{R}^{*2}(2) = 0.421$$

$$R^2_{\min} = R^2_{\min}(2) = 0.323$$



Шаг 3. Выбираем третью переменную, включаемую в модель:

$$R^2_{y/x4,x3,x1} = 0.485 \qquad R^2_{y/x4,x3,x2} = 0.484 \qquad R^2_{y/x4,x3,x5} = 0.498$$

$$\max_{i=\overline{1,2,5}} \bar{R}^2_{y/x_4,x_3,x_i} = \bar{R}^2_{y/x_4,x_3,x_5} = 0{,}498$$

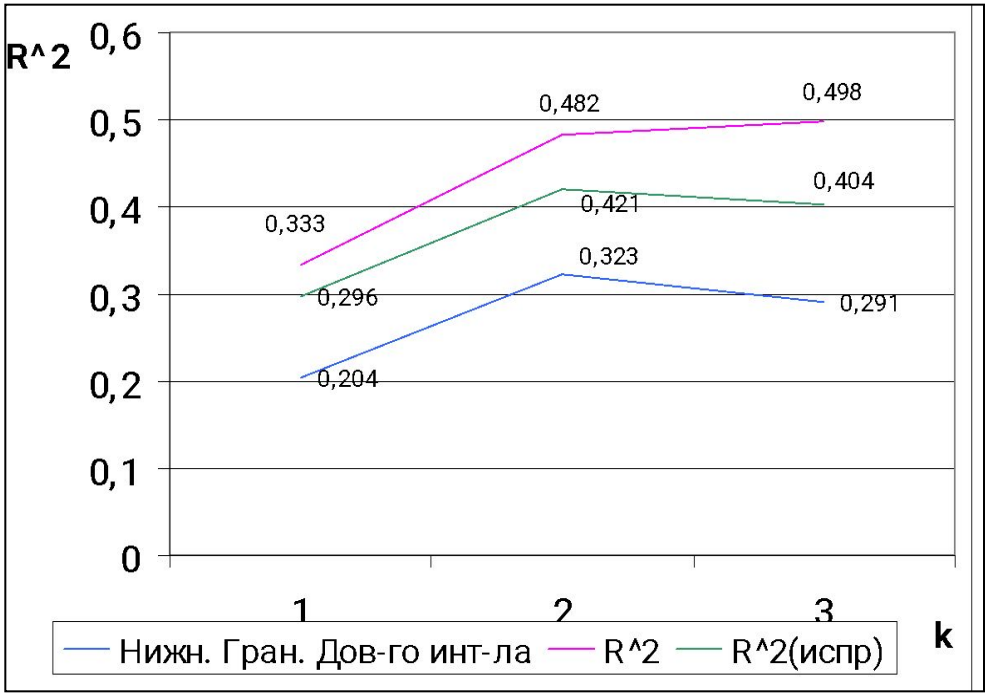
$x^{(3)}$ -Число орудий поверхностной обработки почвы на 100 га;

$x^{(4)}$ - Количество удобрений, расходуемых на гектар (т/га);

$x^{(5)}$ -Количество средств химической защиты растений, расходуемых на гектар (ц/га);

$$\bar{R}^{*2}_{y/x4,x3,x5} = \bar{R}^{*2}(3) = 0.404$$

$$R^2_{\min} = R^2_{\min}(3) = 0.291$$



$$R^2_{\min}(3) < R^2_{\min}(2)$$

$k_0 = 2$ - существенные объясняющие переменные x3, x4:

Regression Summary for Dependent Variable: Y (Spreadsheet1
R= ,69455721 R?= ,48240972 Adjusted R?= ,42151675
F(2,17)=7,9223 p<,00371 Std.Error of estimate: 1,5026

Beta	Std.Err. of Beta	B	Std.Err. of B	t(17)	p-level
		7,291770	0,656561	11,10601	0,000000
0,386342	0,174569	0,281810	0,127336	2,21312	0,040856
0,565605	0,174569	3,473494	1,072065	3,24000	0,004814

$$\hat{y} = \underset{(0,66)}{7,29} + \underset{(0,18)}{0,28} \cdot x_3 + \underset{(1,07)}{3,47} \cdot x_4$$

$x^{(3)}$ -Число орудий поверхностной обработки почвы на 100 га;

$x^{(4)}$ - Количество удобрений, расходуемых на гектар (т/га);

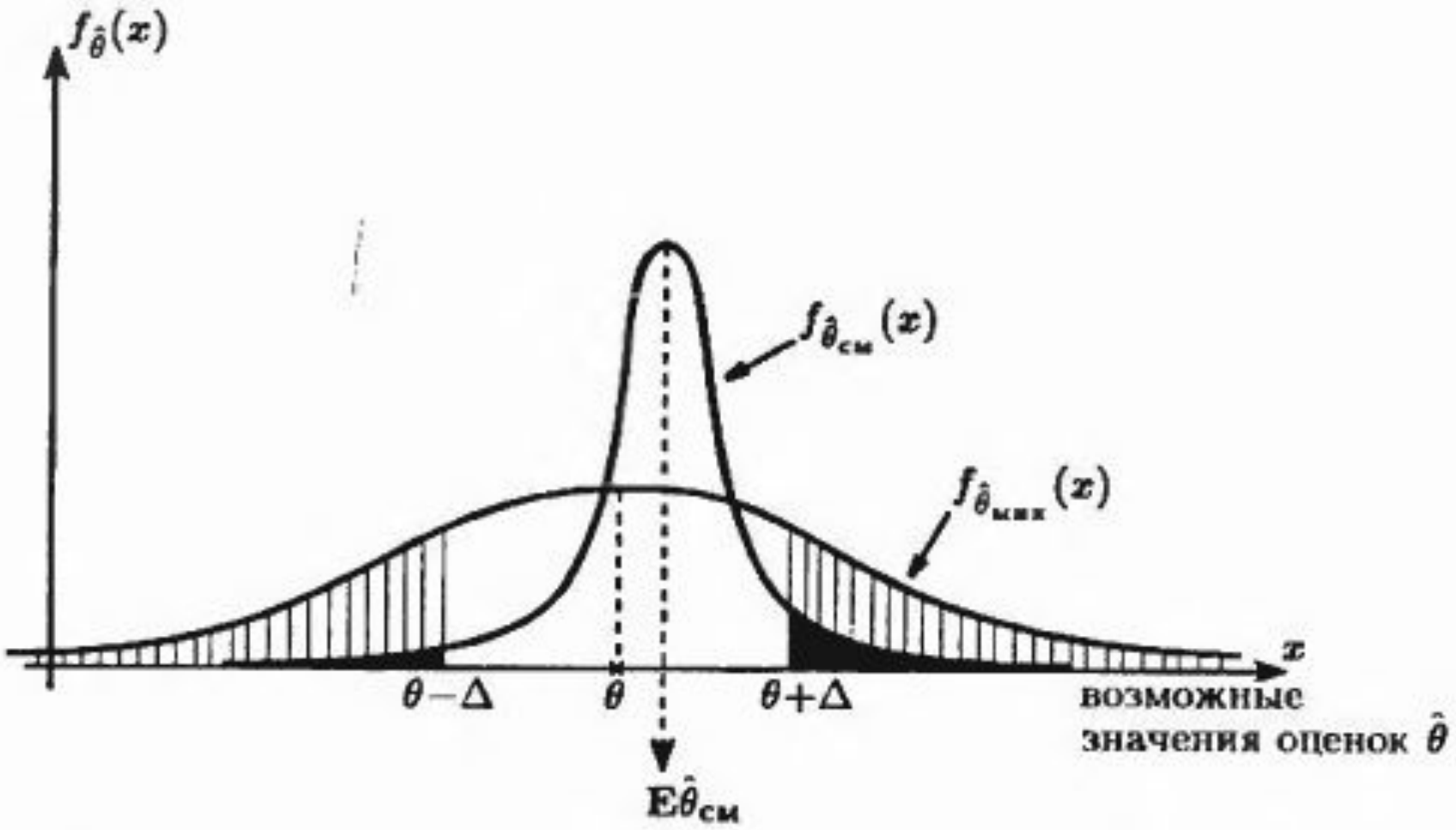
Y – Урожайность зерновых культур (ц/га);

Метод ридж-регрессии

Устранение мультиколлинеарности путем построения смещенных оценок (**ридж-регрессия** или **«гребневая регрессия»**).

$$\bar{b}_\tau = (X^T X + \tau \cdot E_{p+1})^{-1} X^T Y$$

Добавление к диагональным элементам матрицы $(X^T X)$ «гребня» τ (τ - некоторое положительное число, $0{,}1 \le \tau \le 0{,}4$, E_{p+1} - единичная матрица $(p+1)$ порядка) с одной стороны, делает получаемые при этом оценки смещенными, а с другой,- превращает матрицу $X^T X$ из «плохо обусловленной» в «хорошо обусловленную».


$f_{\hat{\theta}}(x)$
$f_{\hat{\theta}_{см}}(x)$
$f_{\hat{\theta}_{мнк}}(x)$
x
$\theta-\Delta$
θ
$\theta+\Delta$
возможные
значения оценок $\hat{\theta}$
$E\hat{\theta}_{см}$