Диагностика нетипичных наблюдений

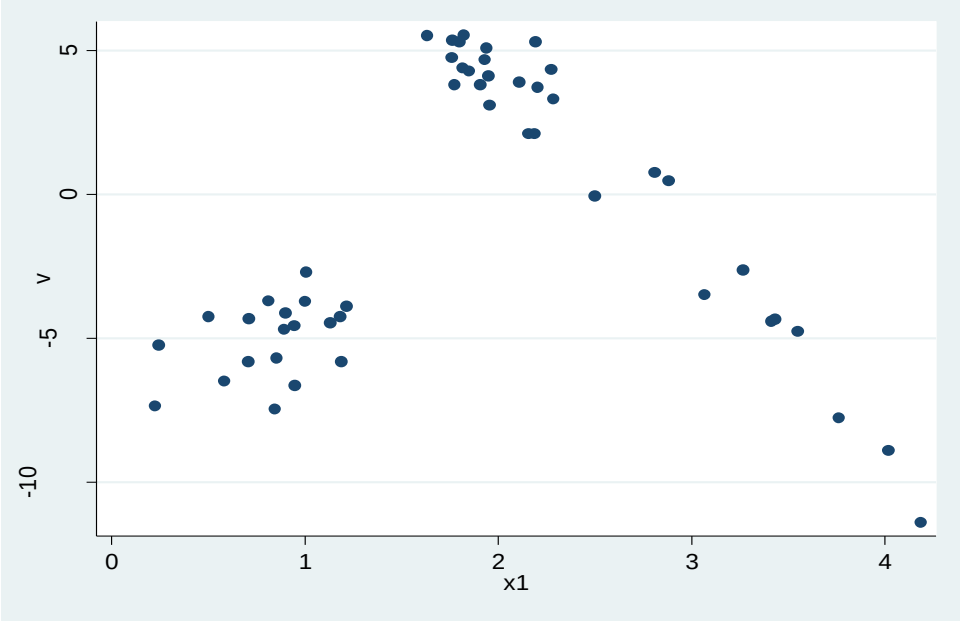
План лекции

- Множественные выбросы
  - Кластеризация
  - Изменение функциональной формы
- Одиночные выбросы
  - Классификация выбросов
  - Коррекция

Множественные выбросы (анализ диаграммы рассеяния)

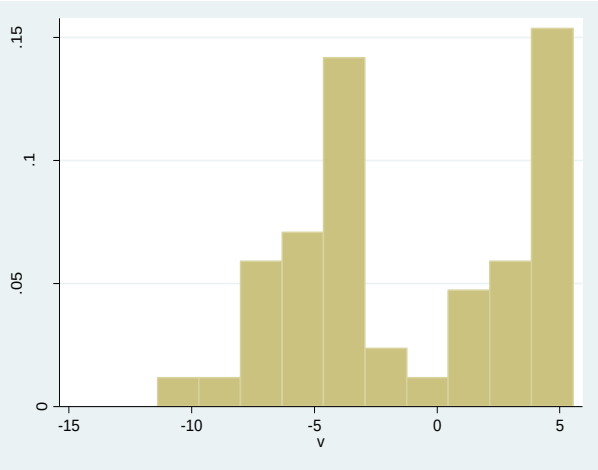
Пусть мы изучаем некую зависимость Y(X) (например, доли расходов на питание в совокупных расходах населения выборки стран от ВВП).

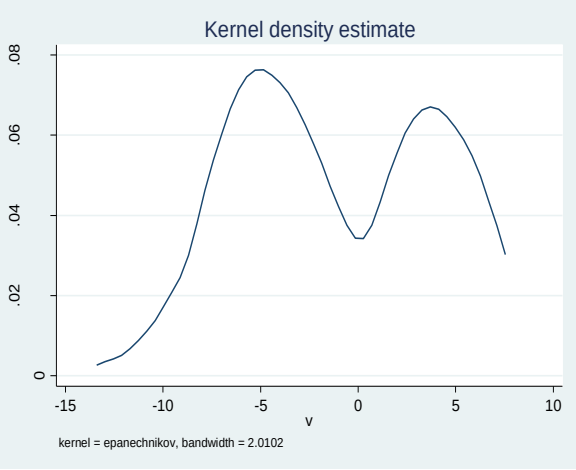
Диаграмма рассеяния говорит о том, что выборка неоднородна.



Множественные выбросы (анализ эмпирических распределений)

Гистограмма и ядерная оценка функции плотности зависимой переменной показывают, что выборка описывается смесью двух распределений

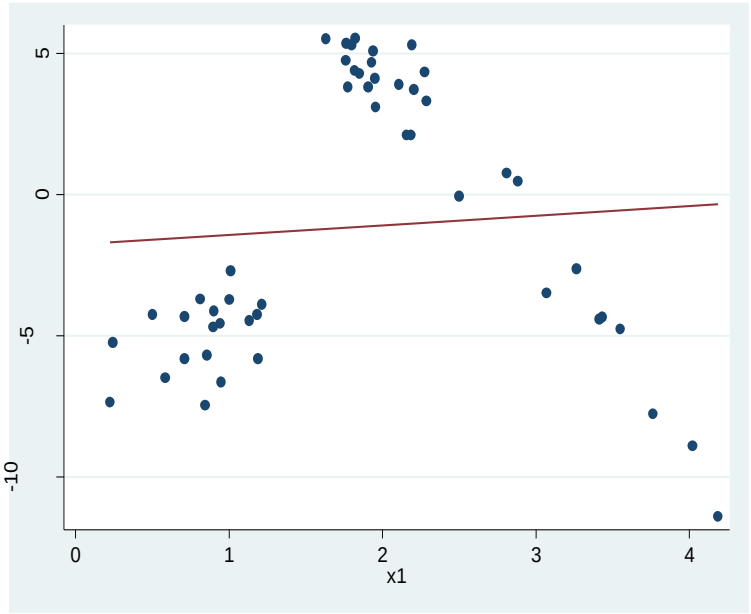




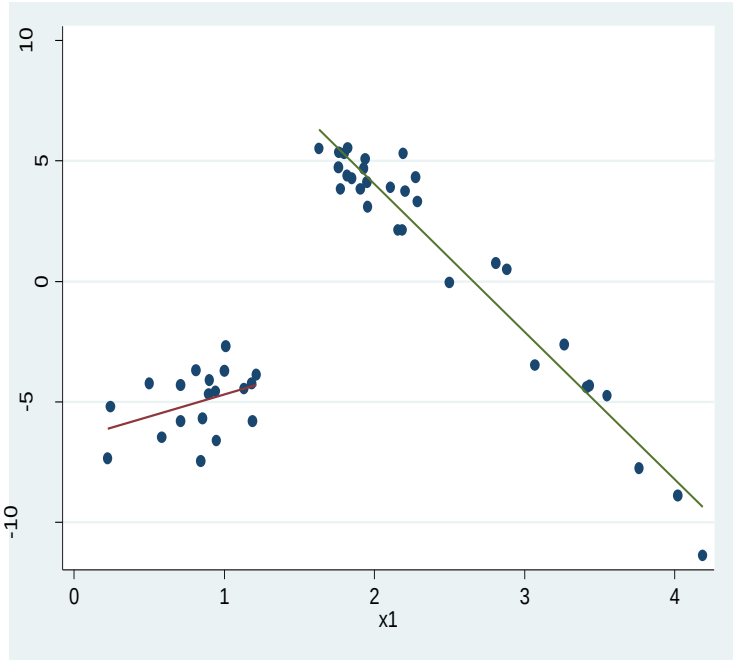
$$F(Y) = (1-\alpha)F_1(Y) + \alpha F_2(Y)$$

Множественные выбросы (оценивание)

Оценка без учета неоднородности



Оценка с учетом неоднородности



Множественные выбросы (учет кластеризации)

Для учета неоднородности во второй модели используются мультипликативные дамми.

Из таблицы видно, что удачная кластеризация радикально повышает качество подгонки

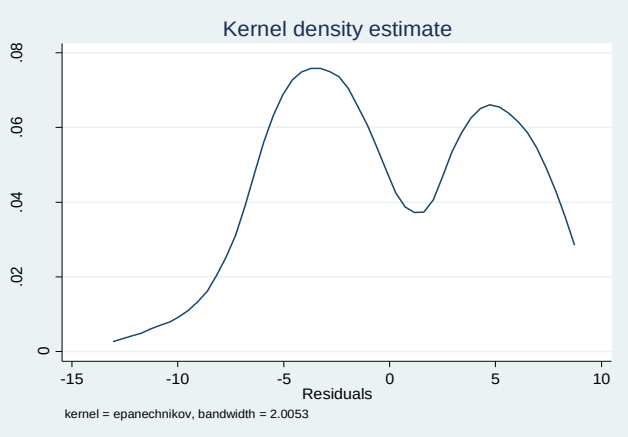
Переменные	Модель 1	Модель 2
X	0.3405	1.8380
D		22.8460***
D*X		-7.9744***
_cons	-1.7647	-6.5375***
N	50	50
r2	**0.0049**	**0.9462**
r2_a	**-0.0158**	**0.9427**

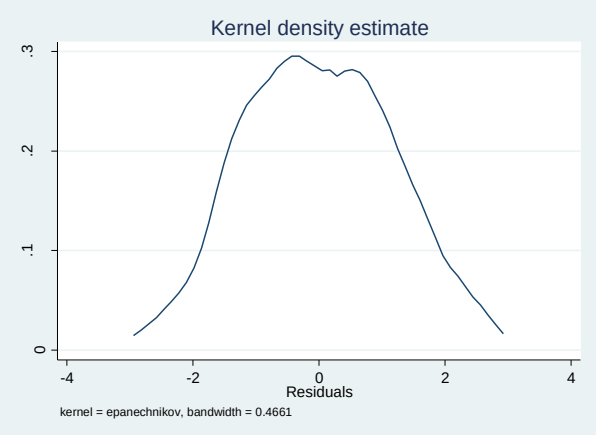
Множественные выбросы (учет кластеризации)

Анализ остатков регрессий показывает, что кластеризация полностью снимает в случае модели 2 проблему множественных выбросов

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
e_m1	50	0.8382	0.0006	9.91	0.0071
e_m2	50	0.8219	0.3623	0.91	0.6333

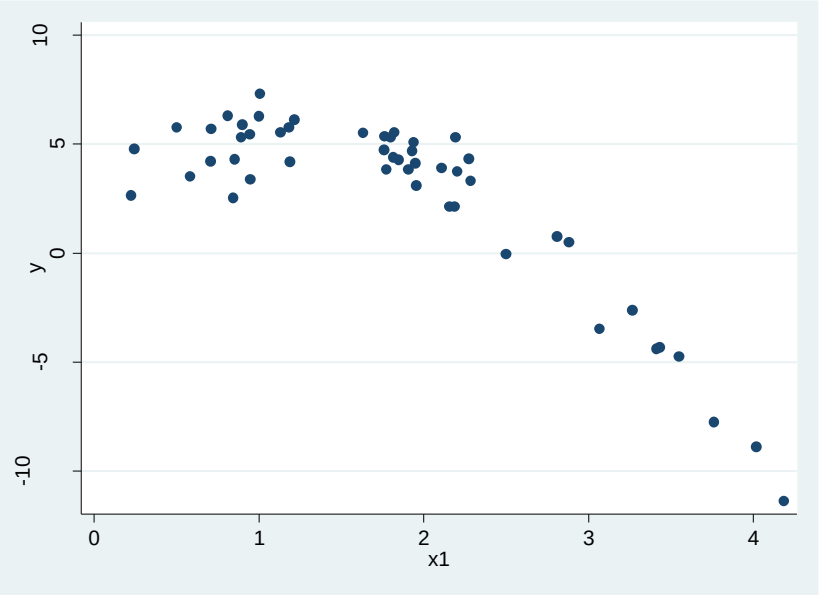




Множественные выбросы (учет нелинейности)

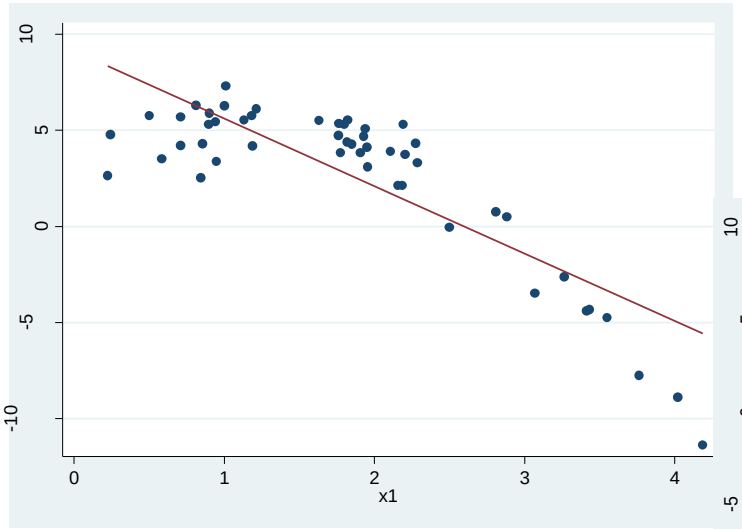
Пусть теперь мы изучаем зависимость Y(X), где Y – продажи, а X – реклама.

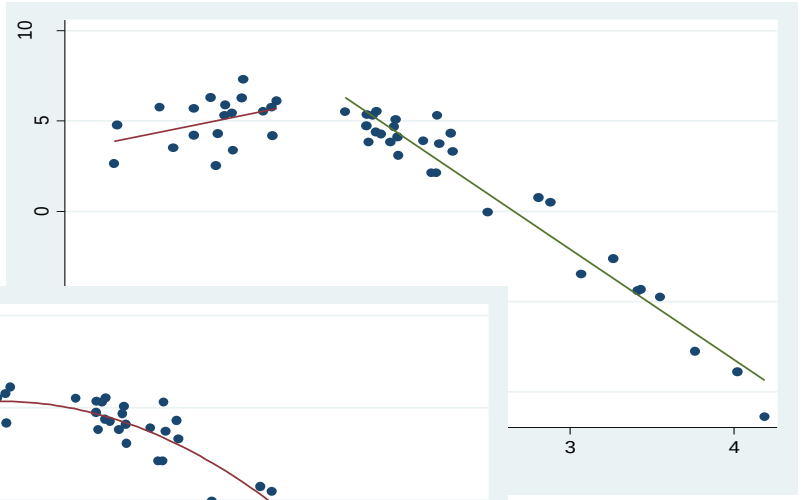
Диаграмма рассеяния опять говорит нам о том, что выборка неоднородна.

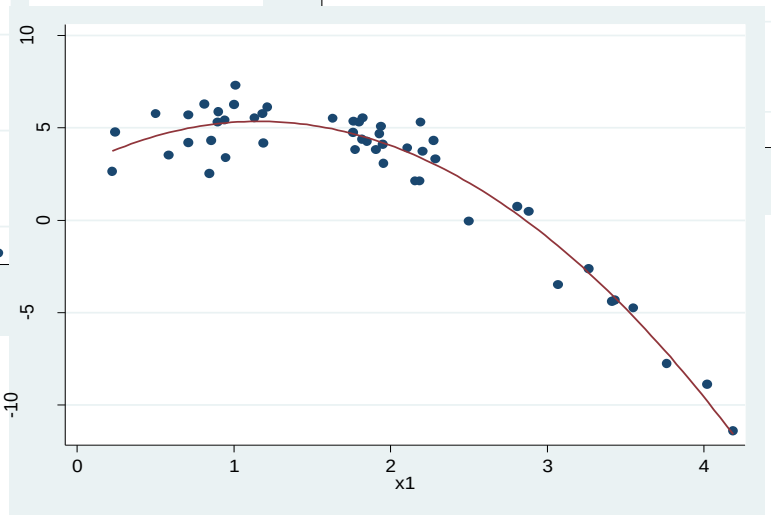


Множественные выбросы (учет нелинейности)

Мы можем проигнорировать неоднородность, можем ввести дамми на кластеры, а можем изменить функциональную форму уравнения модели







Множественные выбросы (учет нелинейности)

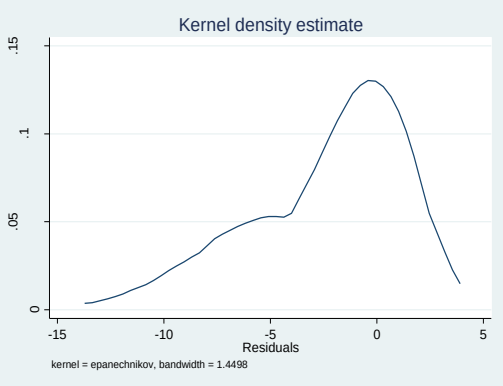
Сопоставим оценки 3-х моделей

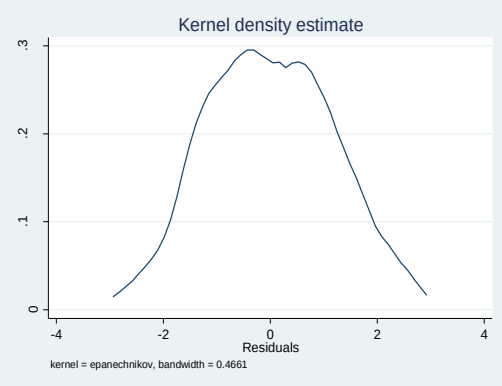
Переменные	Модель 1	Модель 2	Модель 3
x	-3.5135***	1.8380	4.2824***
D		12.8460***	
D*X		-7.9744***	
X*X			-1.8492***
_cons	9.1310***	3.4625***	2.8791***
r2	**0.6961**	**0.9285**	**0.9409**
r2_a	0.6898	0.9238	0.9384

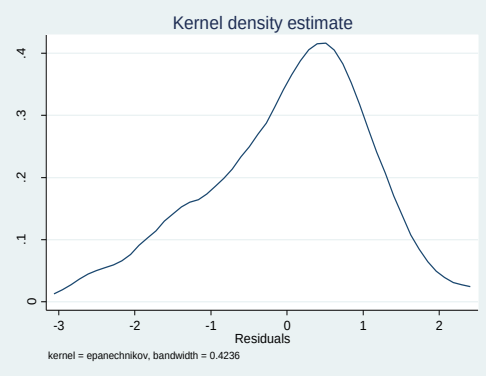
Сравнение по критериям качества подгонки показывает явное преимущество модели 3 с квадратичной зависимостью от X

Множественные выбросы (учет нелинейности)

Анализ остатков тоже не дает веских оснований сомневаться в качестве 3-ей модели







Skewness/Kurtosis tests for Normality

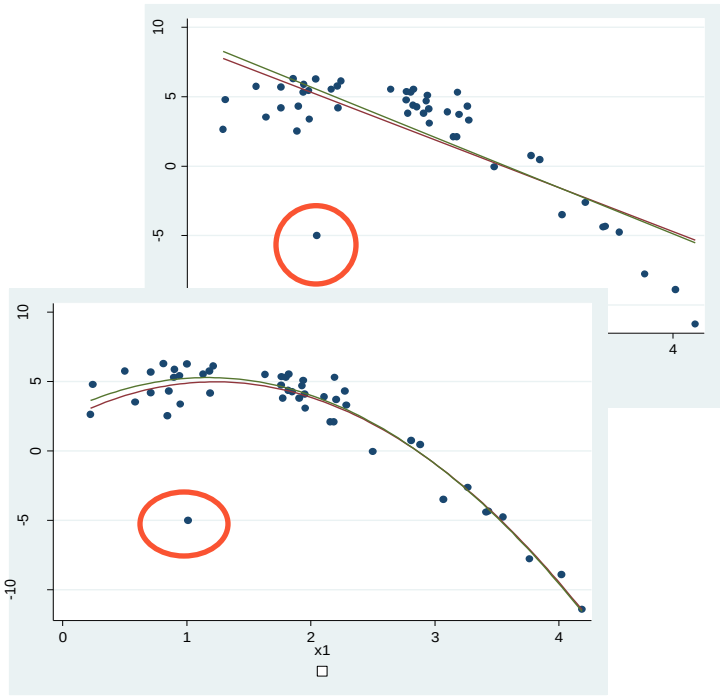
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
e_m1	50	0.0072	0.6422	6.75	0.0342
e_m2	50	0.8219	0.3623	0.91	0.6333
e_m3	50	0.0877	0.7959	3.16	0.2063

Множественные выбросы (вывод)

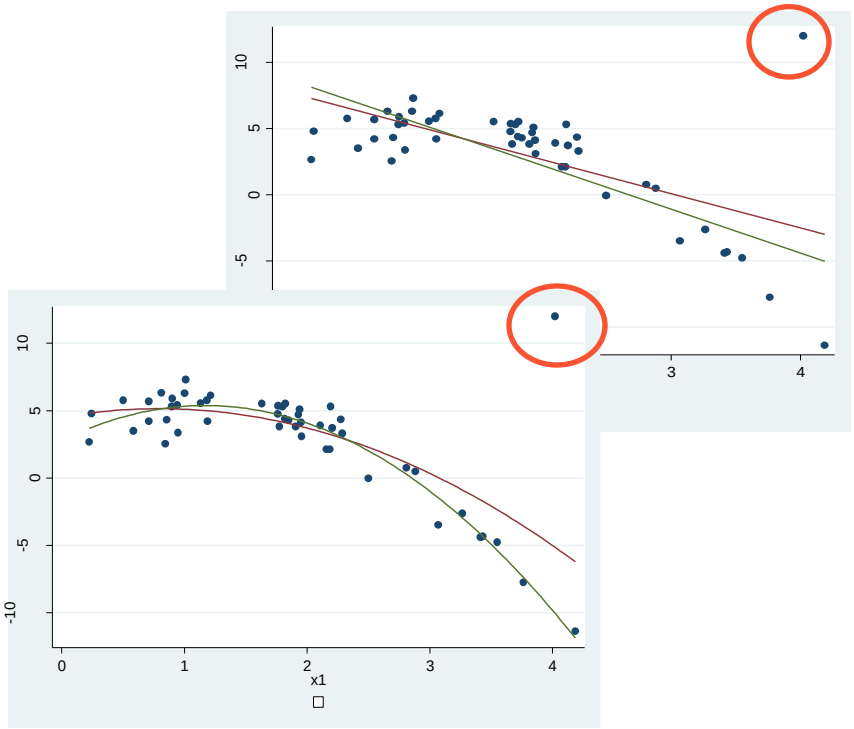
- Нетипичные наблюдения – понятие относительное.
- То, что нетипично для одной модели, может вполне удачно вписываться в другую.
- Если нетипичных наблюдений много, следует искать способ улучшить спецификацию модели.

Одиночные выбросы

В одних случаях наличие выбросов почти не меняет вид зависимости



В других возникают довольно заметные искажения

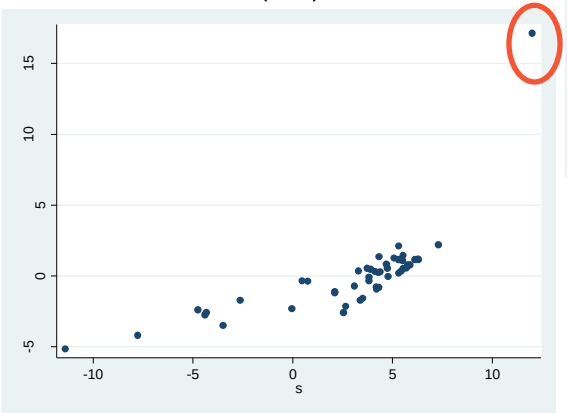


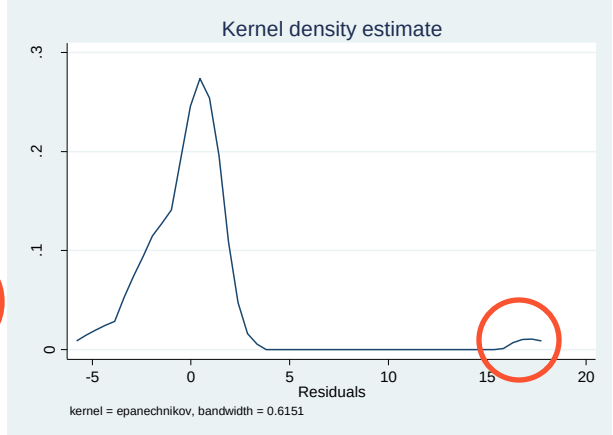
С чем это связано?

Одиночные выбросы (визуальный анализ)

- Рассмотрим пространственную выборку.
- Пусть по графику остатков обнаружен одиночный выброс.

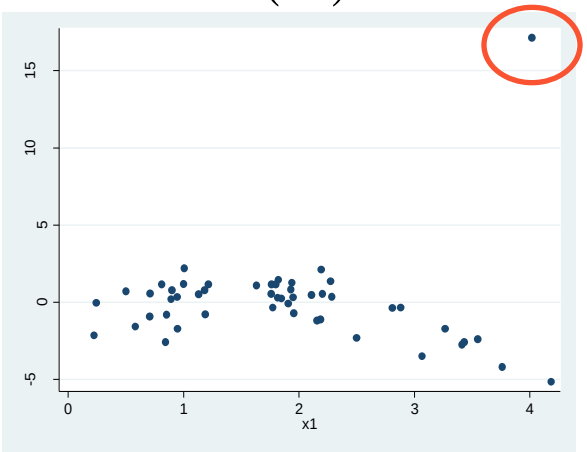
$\hat{\varepsilon}(Y)$





$f(\hat{\varepsilon})$

$\hat{\varepsilon}(X)$



Одиночные выбросы (теоретический анализ)

- Попробуем выяснить, что будет с оценками, если удалить одиночный выброс
- Переупорядочим наблюдения так, чтобы выброс оказался в последней строке

$$X = \begin{pmatrix} \tilde{X} \\ X_n' \end{pmatrix} \quad Y = \begin{pmatrix} \tilde{Y} \\ Y_n' \end{pmatrix} \quad X' = \begin{pmatrix} \tilde{X} & X_n \end{pmatrix}$$

$$X'X = \tilde{X}'\tilde{X} + X_n X_n' \qquad X'Y = \tilde{X}'\tilde{Y} + X_n Y_n$$

$$\tilde{X}'\tilde{X}\beta_{(-n)} = \tilde{X}'\tilde{Y}$$

- $\beta_{(-n)}$ вектор коэффициентов регрессии при удалении выброса

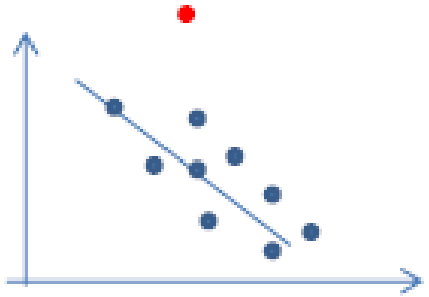
Связь коэффициентов до и после удаления выбросов

- В результате цепочки выкладок получаем

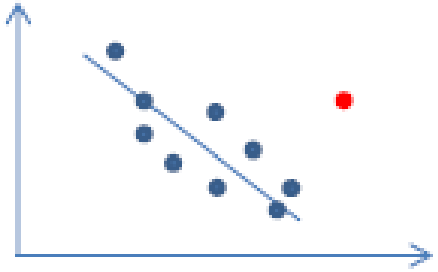
$$\hat{\beta}_{(-n)} = \hat{\beta} - \frac{(X'X)^{-1} X_n}{1 - P_{nn}} \hat{\varepsilon}_n$$

- Овалом выделено смещение оценки от удаления выброса
- Выводы:
  - Если остаток $\hat{\varepsilon}_n$ велик, оценки могут сильно измениться
  - Если $P_{nn} \approx 1$, то после удаления выброса вообще нельзя будет получить адекватные оценки из-за мультиколлинеарности

Классификация выбросов

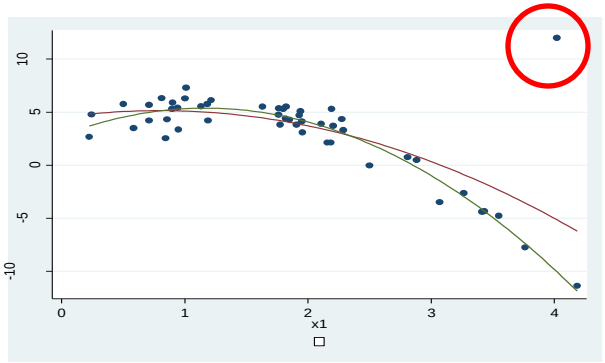


- вертикальный выброс
(остаток $\hat{\varepsilon}_i$ велик)



- точка разбалансировки или выброс с плохим плечом (bad leverage outlier)
($P_{ii} \approx 1$)

Что делать с единичным выбросом?



- Если, как в рассмотренном примере, он одновременно и вертикальный, и с плохим плечом, и при этом его удаление только улучшает результат, его естественно удалить.
- Можно параллельно проводить оценивание с выбросом и без для контроля робастности результатов.
- Ввести дамми на выброс и проверить ее значимость.
- Использовать робастные методы оценивания (медианную регрессию).

Спасибо за внимание!