Кросс-валидация для выявления и предотвращения overfitting

Курс «Искусственные нейронные сети»

Аббакумов В.Л.

Источники

- **Данная презентация в основном является переводом презентации**
- **Andrew W. Moore**

- **Исходный текст (на английском) можно скачать по адресу**
- **http://www.cs.cmu.edu/~awm/tutorials**

О термине overfitting

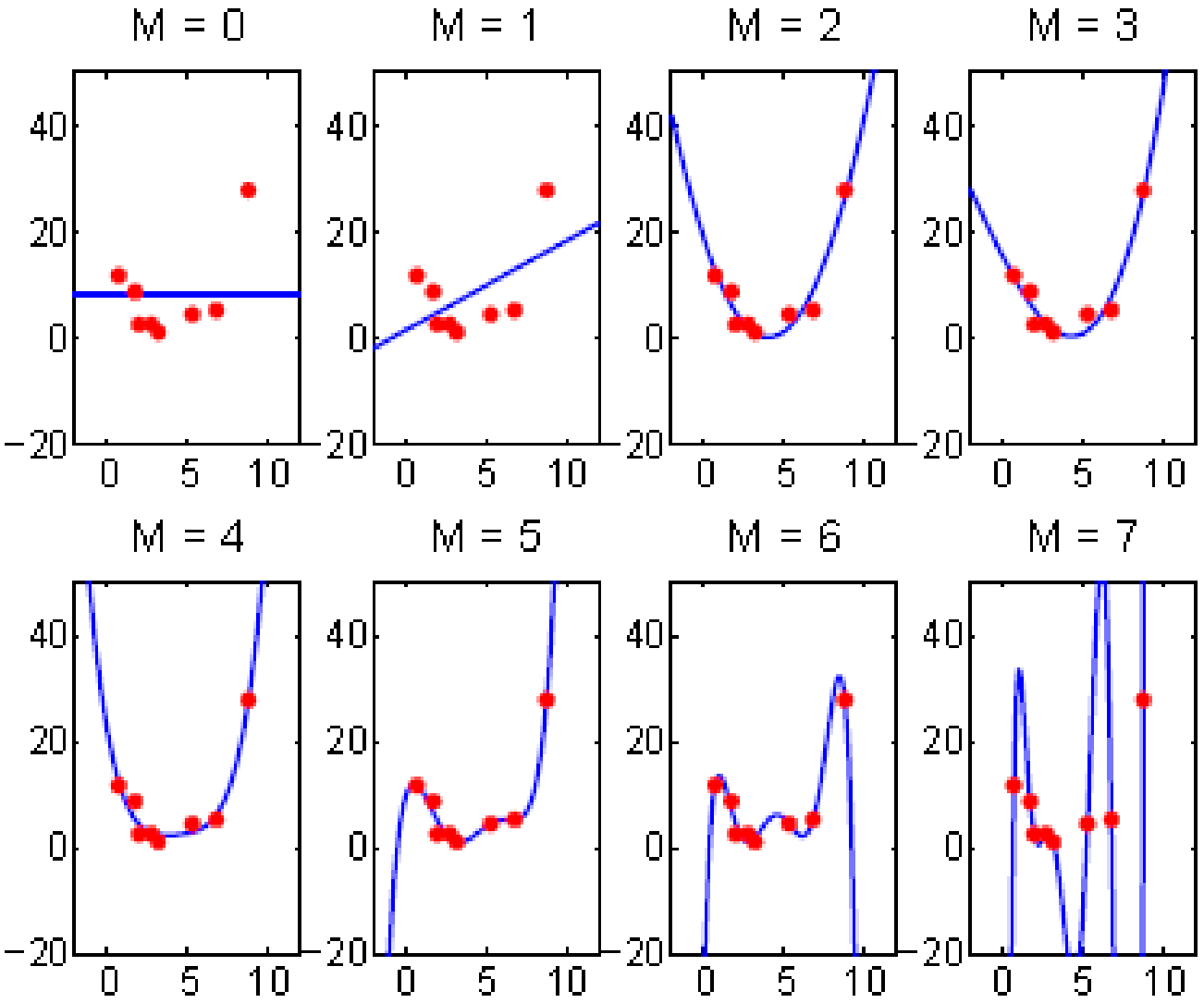
- **Мне неизвестен удачный перевод этого термина…**
- **Наиболее точным представляется «чрезмерная подгонка»**
- **В данной презентации под overfitting’ом будет пониматься**

- **Использование чрезмерно сложной модели для описания данных**

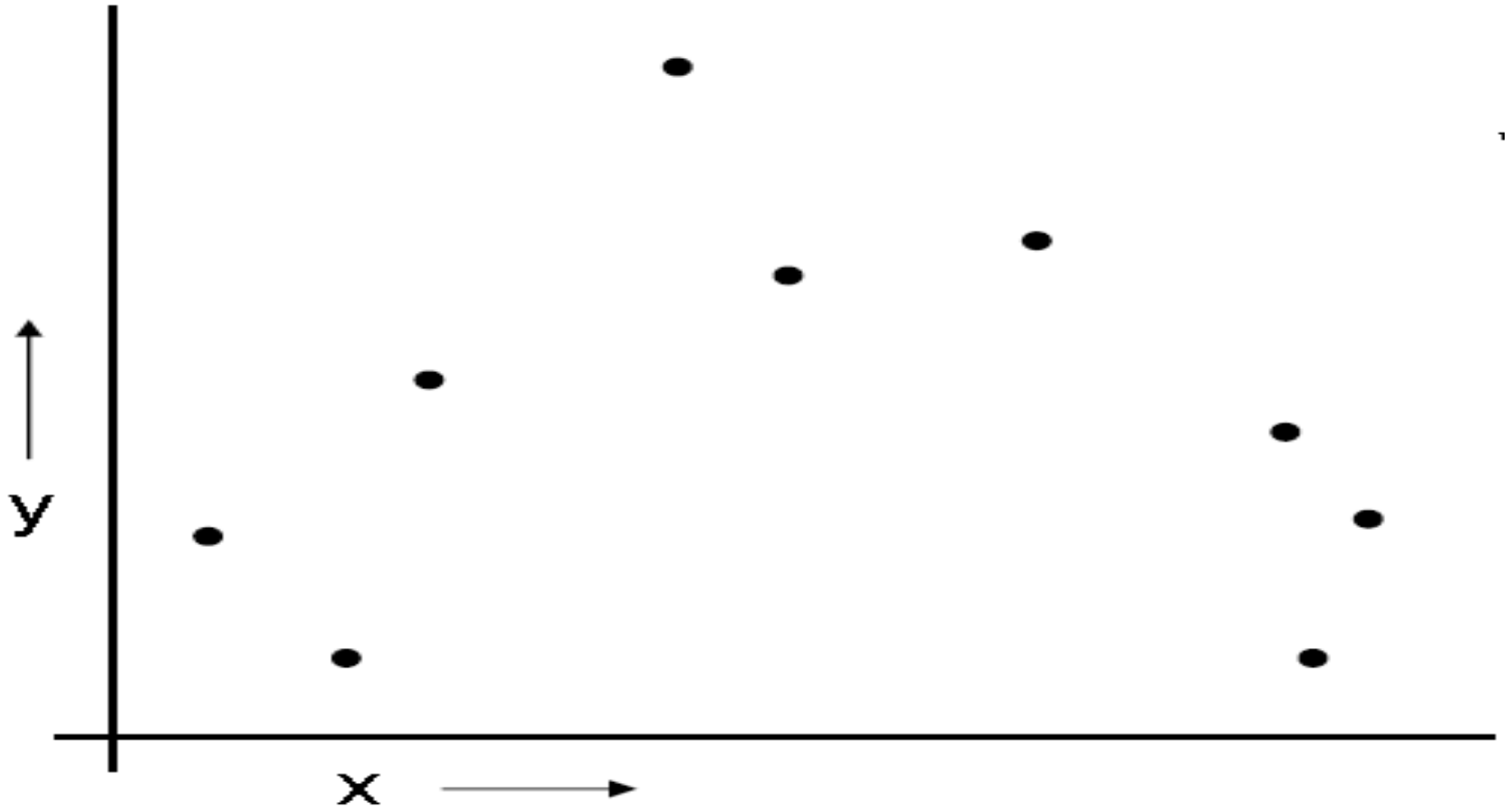
Рассмотрим задачу регрессии

- **Y=f(X)+шум**

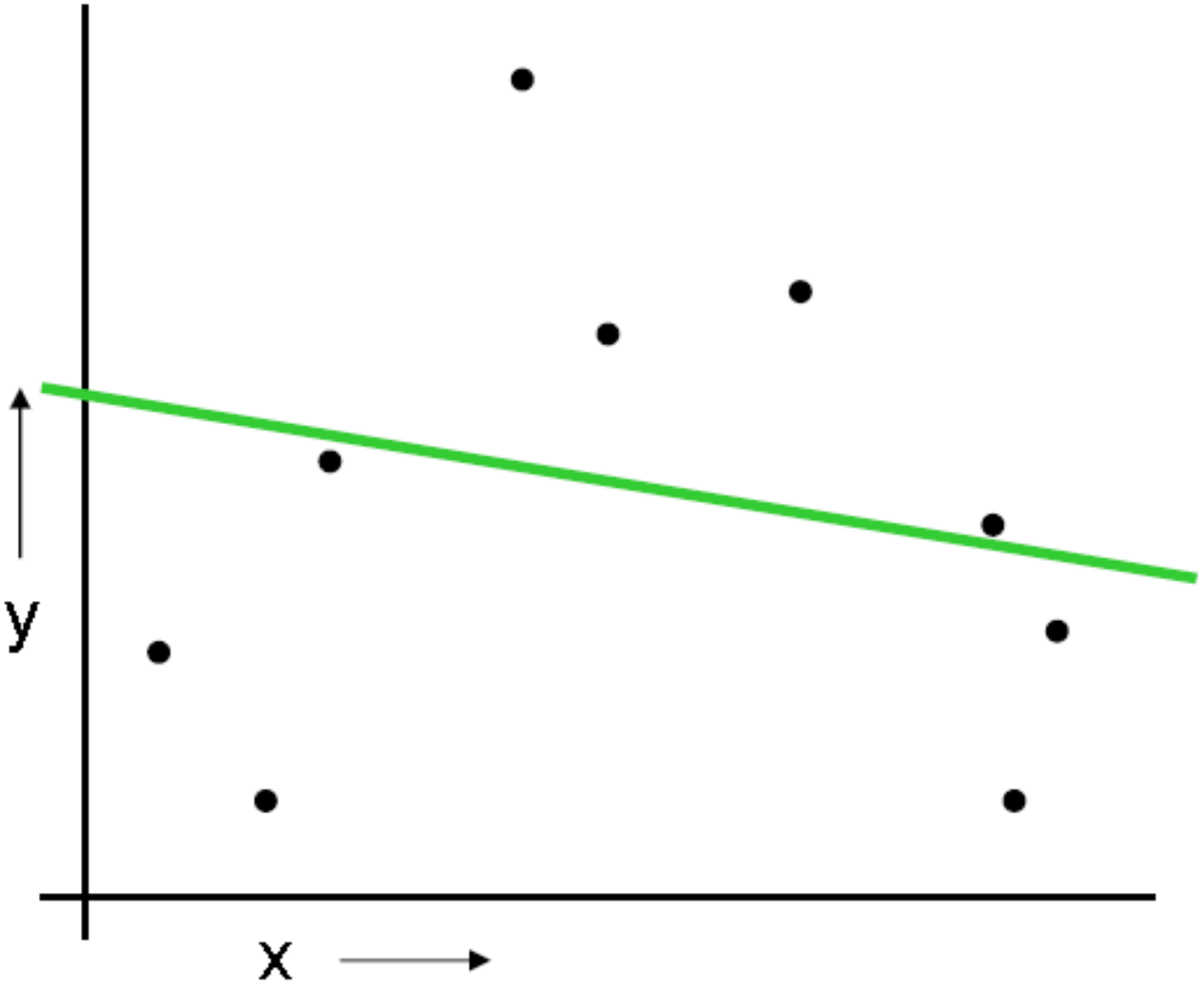
Подгонка многочленами разных степеней – выберите модель



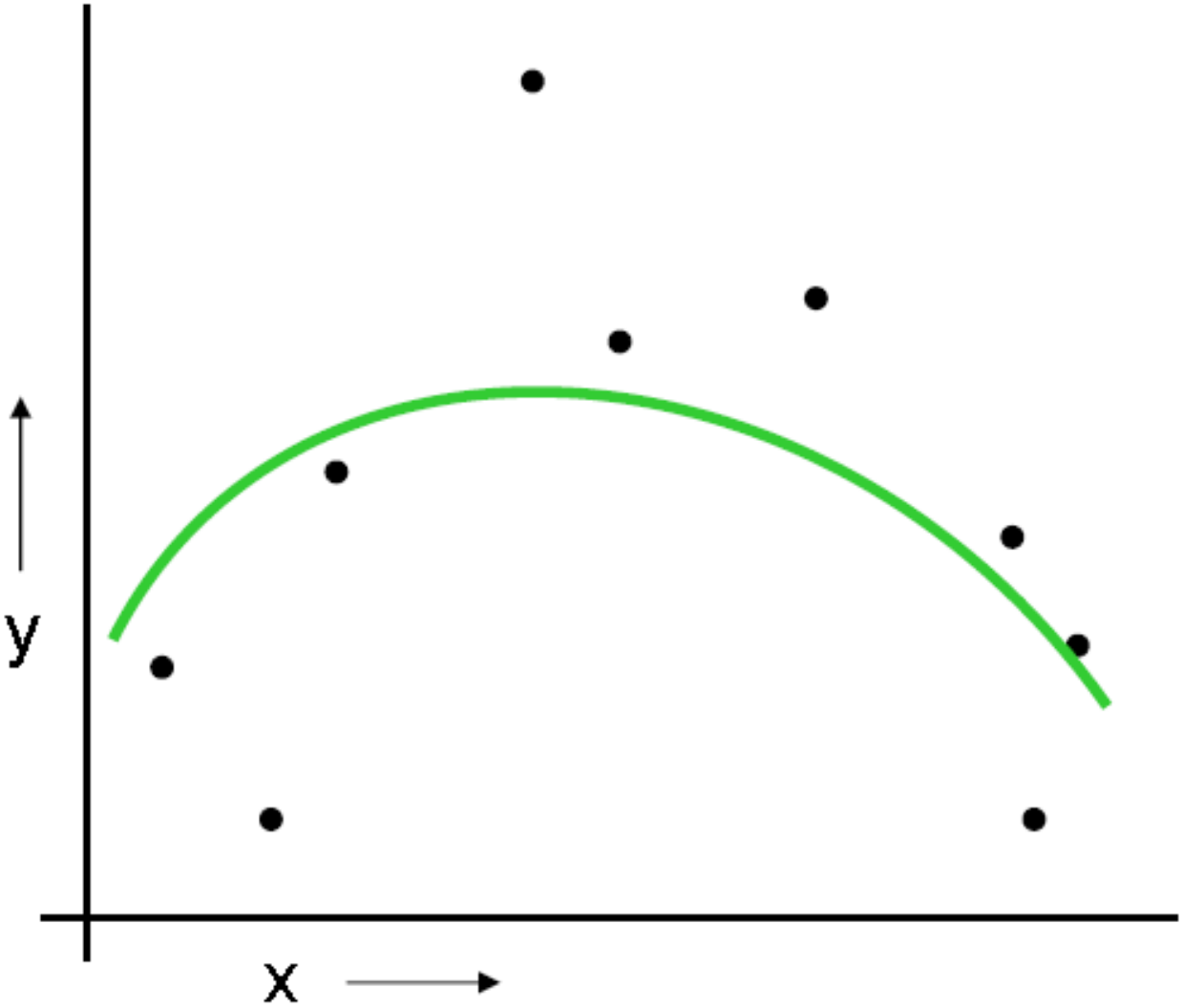
Рассмотрим следующие данные



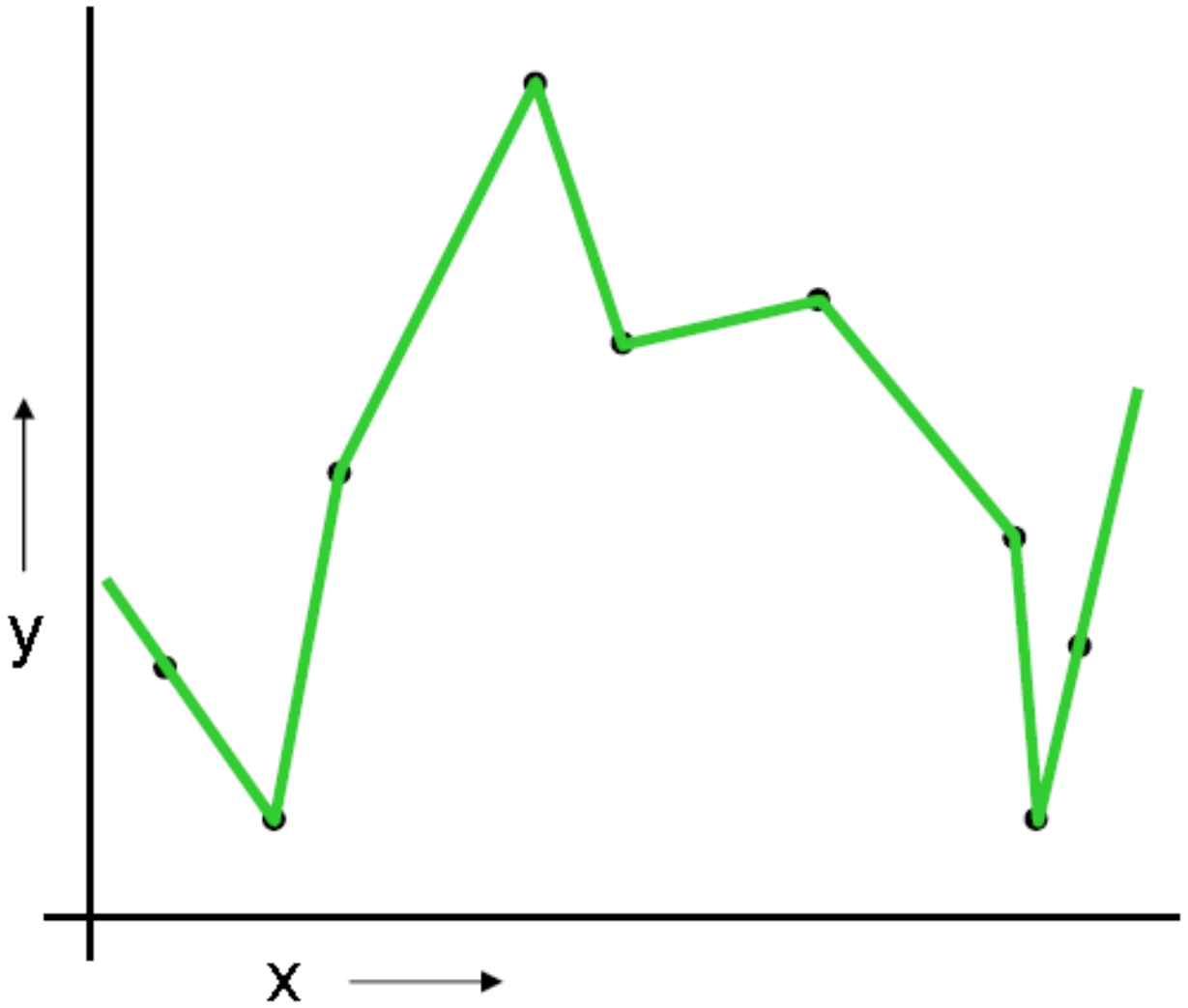
Модель 1: линейная регрессия



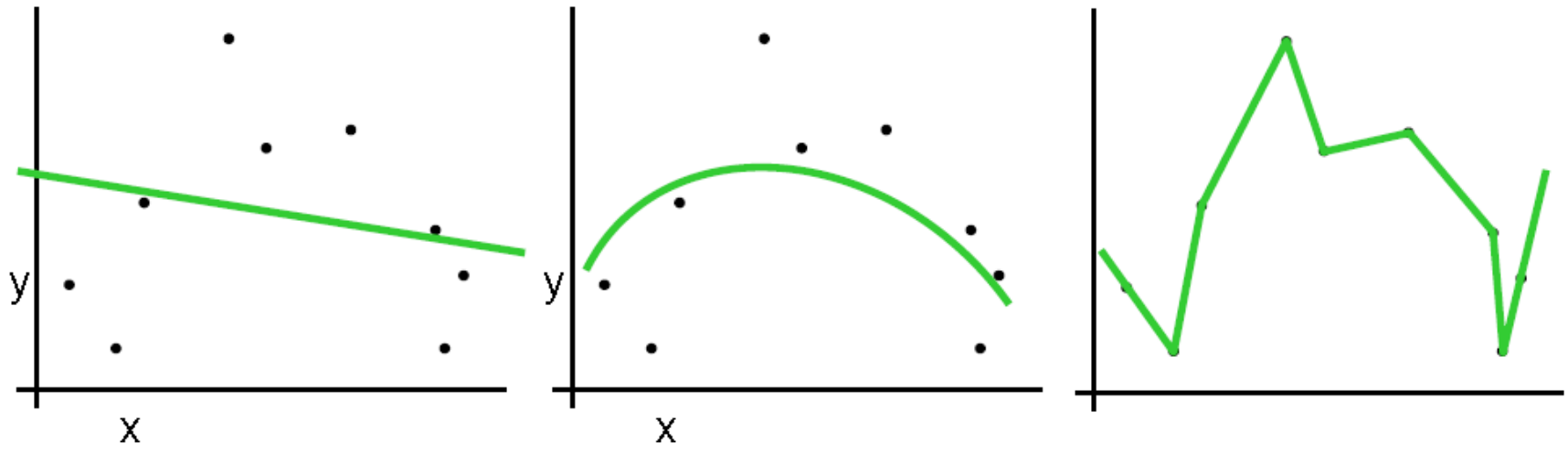
Модель 2: квадратичная регрессия



Модель 3: линейный сплайн



Какую модель выбрать?



Идея 1: «научный» подход

- **Надо выбрать ту из них, которая наилучшим образом подгоняет данные.**
- **То есть ту, у которой наименьшая средняя ошибка.**
- **Или наименьшая средняя квадратичная ошибка.**

Критерий качества

- **Сумма модулей ошибок или**
- **Сумма квадратов ошибок**

В этом случае

- **Лучшей будет сплайн**
- **На втором месте – квадратичная модель**
- **На третьем месте – линейная модель**

Уточняющий вопрос

- **Почему линейная модель не может подгонять данные лучше, чем квадратичная?**

Проблема

- **Используются одни и те же данные**
- **Сначала для подгонки модели**
- **(то есть для вычисления параметров модели)**
- **Потом для оценки качества модели.**

Например

- **Портной научился хорошо шить костюмы для Смита.**
- **Пока он шьет для Смита – все хорошо.**
- **Но если он будет шить для Джонса по тем же лекалам, что и для Смита, результат может быть намного хуже.**

Идея 2: что будет «потом»?

- **Зачем нужна модель?**
- **Чтобы успешно предсказывать будущие наблюдения.**
- **Чтобы для нового значения x предсказать значение Y с маленькой погрешностью.**

- **Наилучшей будет та модель, которая лучше всех будет предсказывать.**
- **То есть описывать (подгонять) новые данные.**

Проблема

- **У нас нет будущих значений…**
- **Они появятся потом…**
- **И будет уже поздно…**

Проблема решаема

- **У нас нет будущих значений…**
- **Так сделаем их из прошлых!**
- **Отберем часть наблюдений и объявим их «будущими»**

Метод тестового множества

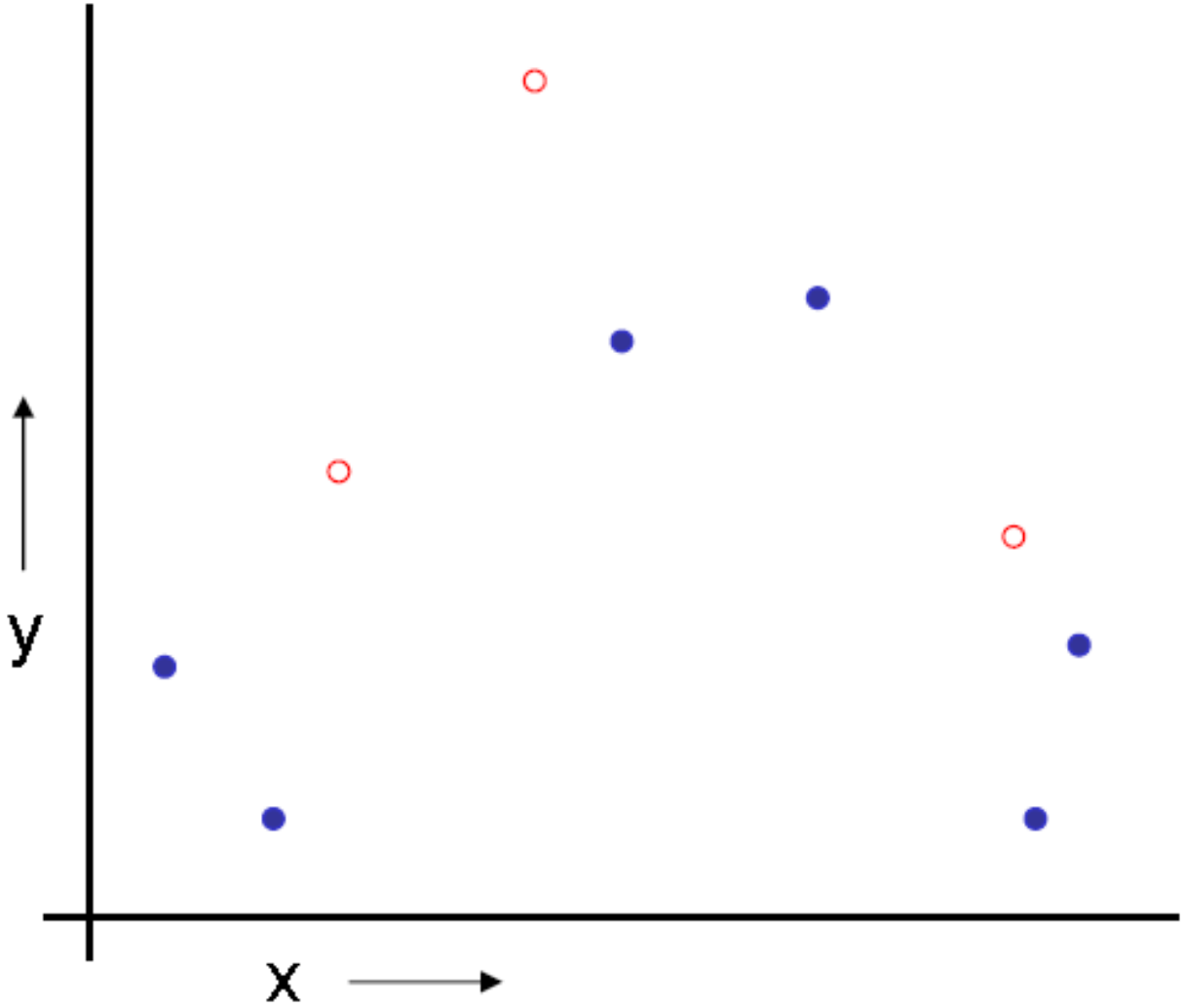
- **Случайным образом выберем 30% всех наблюдений и назовем их «тестовой выборкой».**
- **Остальные 70% наблюдений назовем «обучающей выборкой».**

Интерпретация

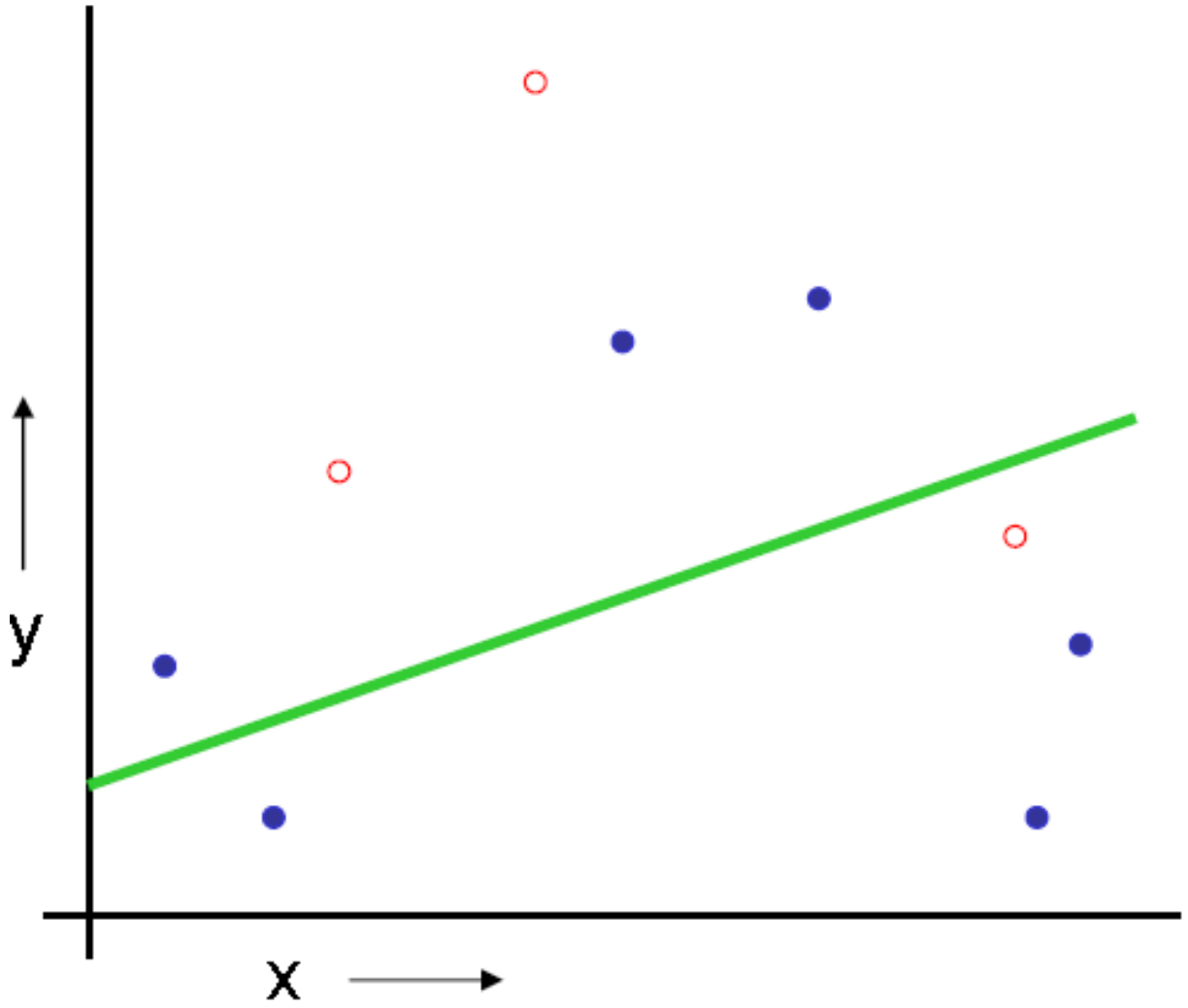
- **Обучающая выборка – прошлое**
- **Тестовая выборка – будущие наблюдения.**

- **По наблюдениям из обучающей выборки построим модель.**
- **Проверим модель на тестовой выборке.**

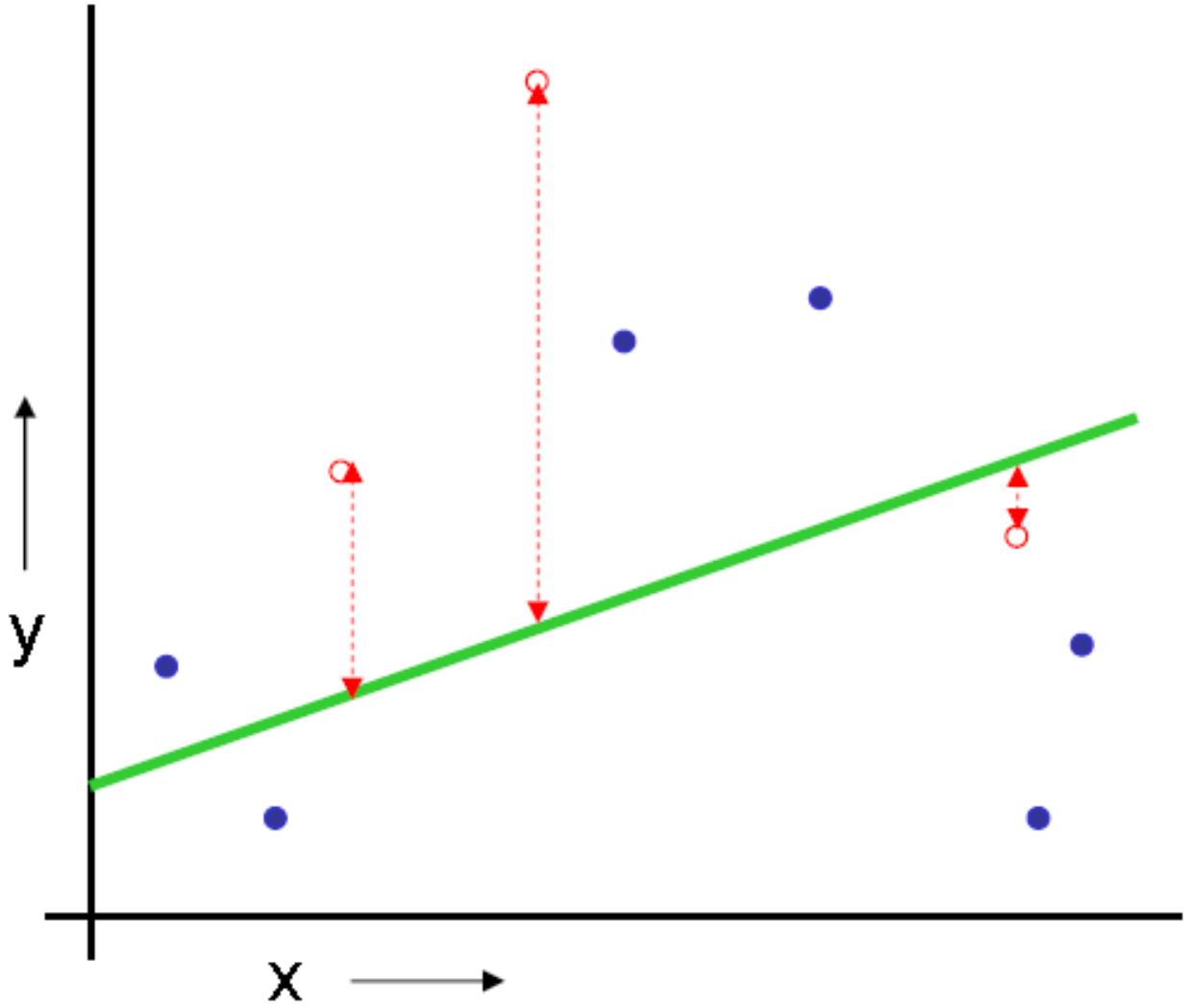
Обучающая и тестовая выборки



Линейная регрессия по обучающей выборке

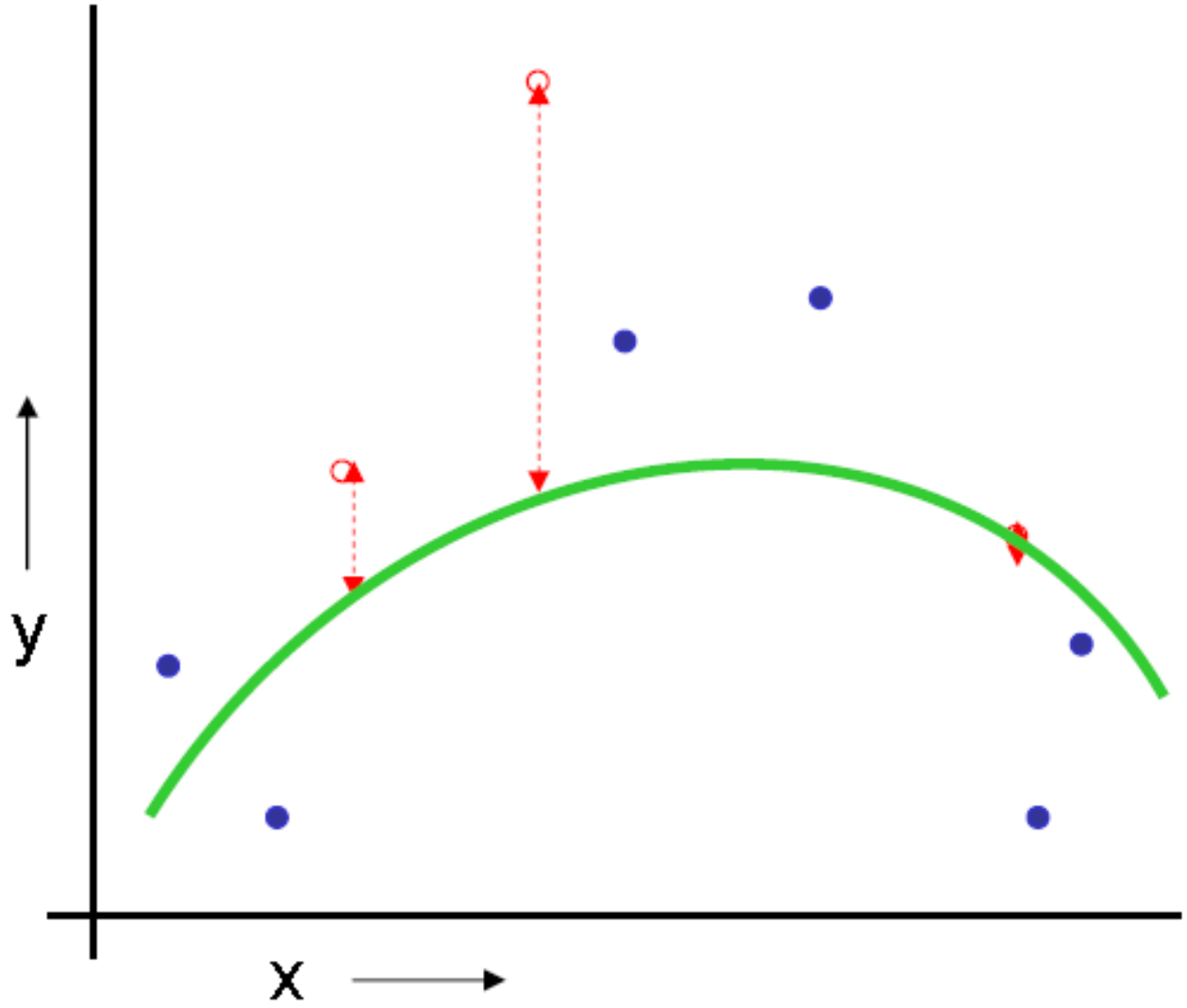


Найдем ошибки на тестовом множестве



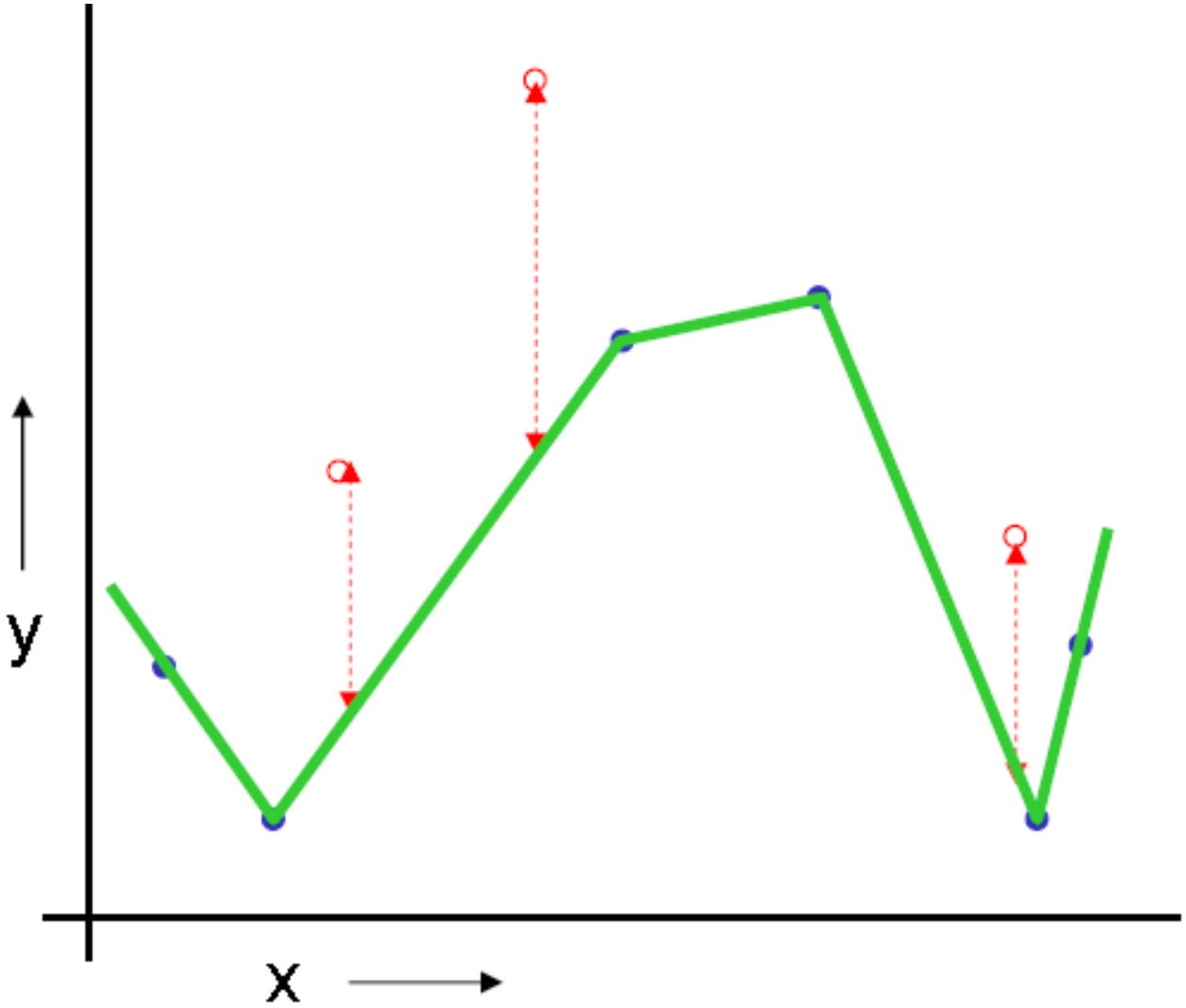
- **Сумма квадратов ошибок на тестовом множестве равна 2.4**

Квадратичная регрессия



- **Сумма квадратов ошибок на тестовом множестве равна 0.9**

Линейный сплайн



- **Сумма квадратов ошибок на тестовом множестве равна 2.2**

Обсуждение метода тестового множества

- **Достоинства**
- **- очень просто реализуется;**
- **- понятен.**

Обсуждение метода тестового множества

- **Недостатки**
- **Расточителен: при построении модели отбрасывается 30% данных**
- **Если данных мало, как распределятся точки между обучающей и тестовой выборками? Дело случая. А это влияет на результат оценивания качества метода.**

- **Другими словами.**
- **Оценка качества модели с помощью тестового множества имеет большую дисперсию**

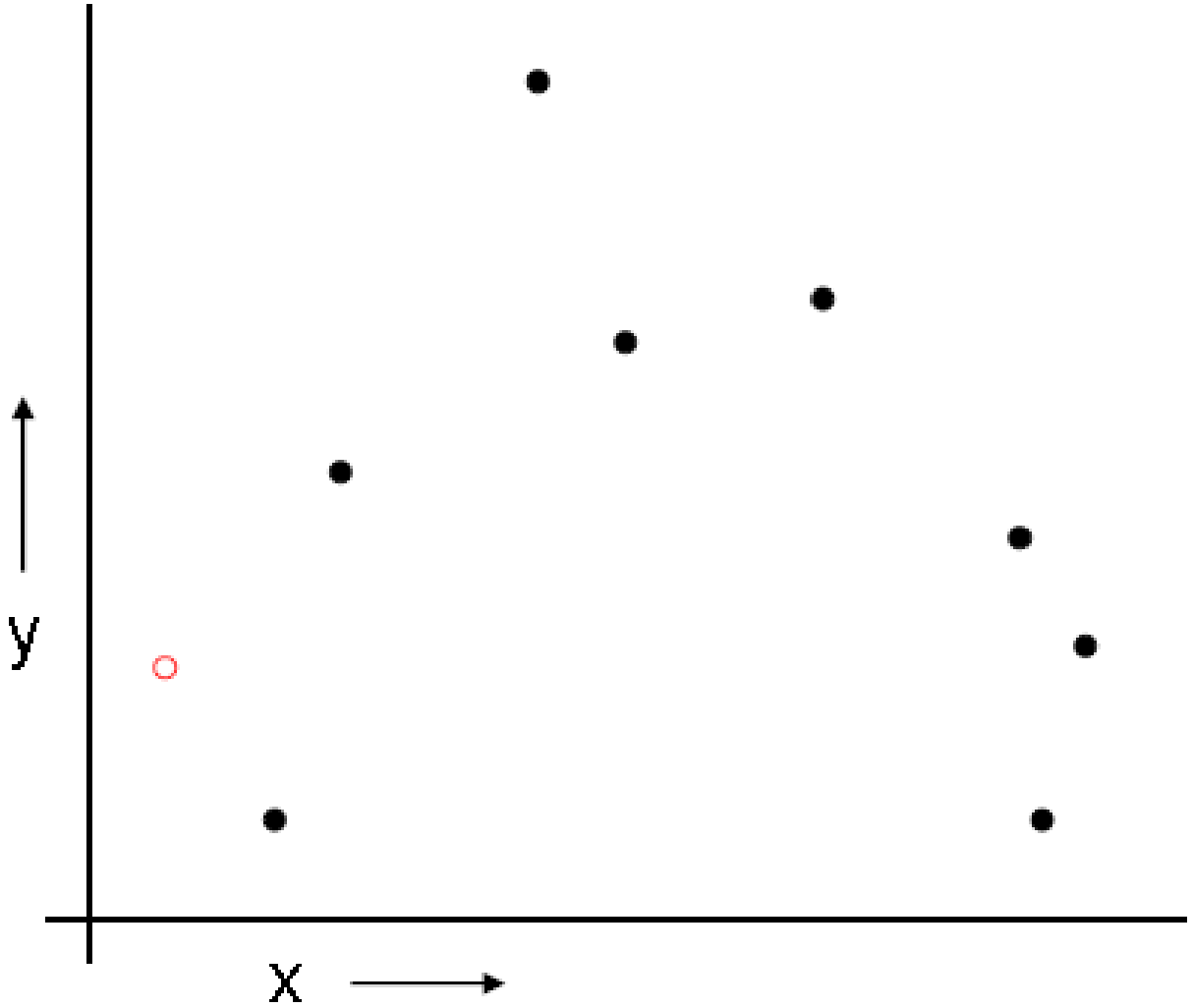
Метод валидации посредством исключенных наблюдений (leave one out validation)

- **Пусть у нас имеется n точек.**
- **Предыдущую процедуру проводим n раз.**
- **Каждый раз тестовое множество состоит из одной точки, каждый раз новой.**
- **За n шагов перебираем все точки множества.**

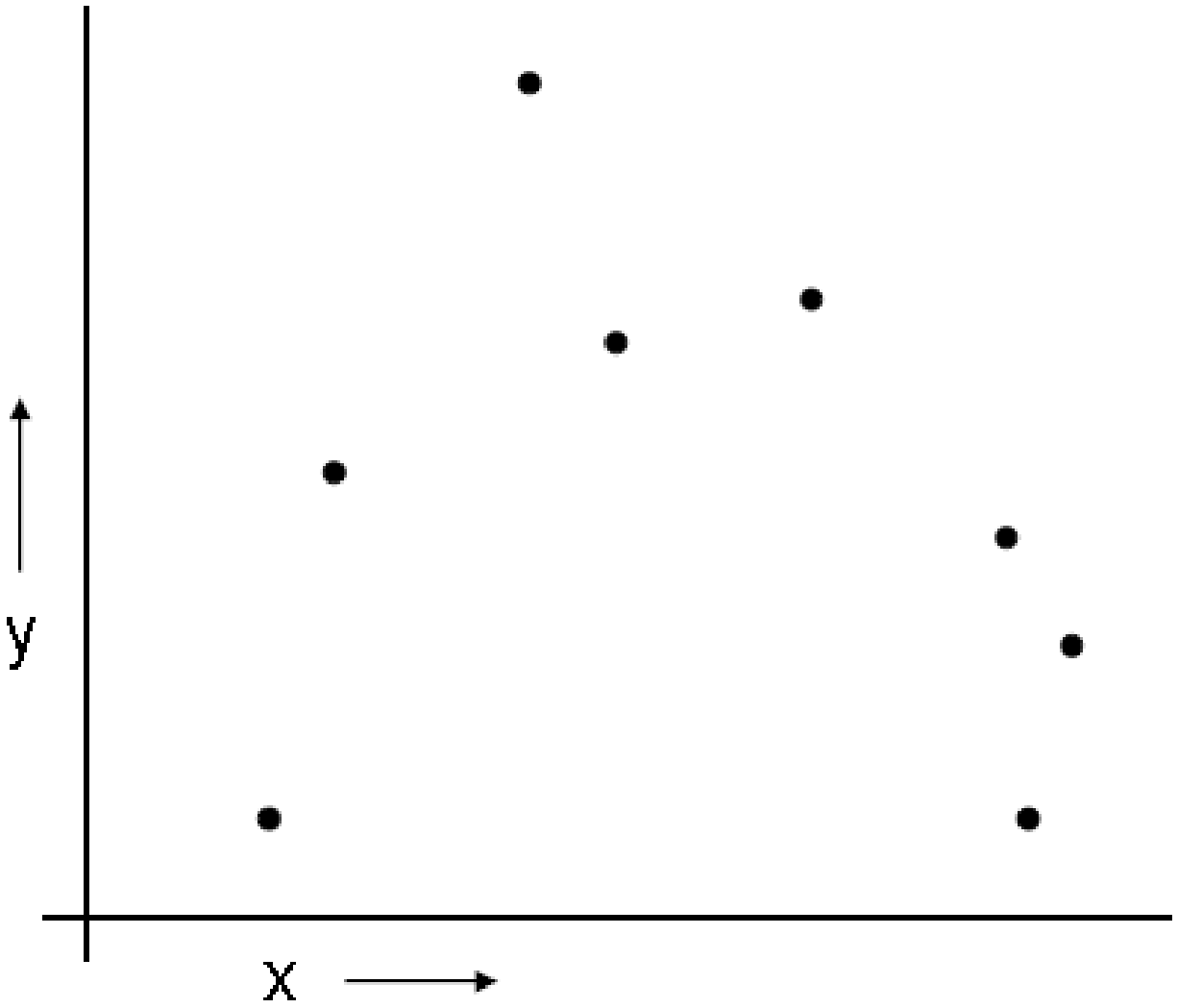
- **Посчитаем среднее значение квадратов ошибок методом валидации посредством исключенных наблюдений.**

- **Проверим линейную регрессионную модель**

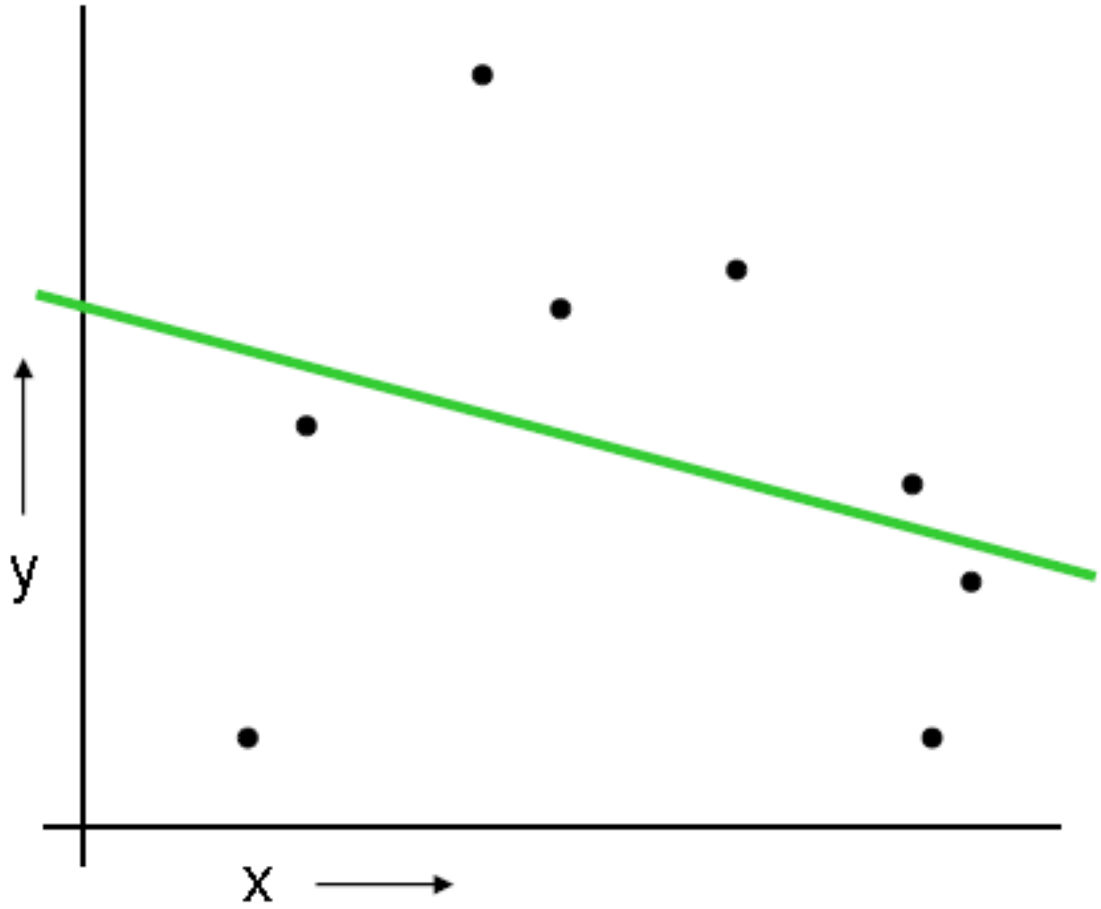
Сейчас тестовая выборка – красная точка



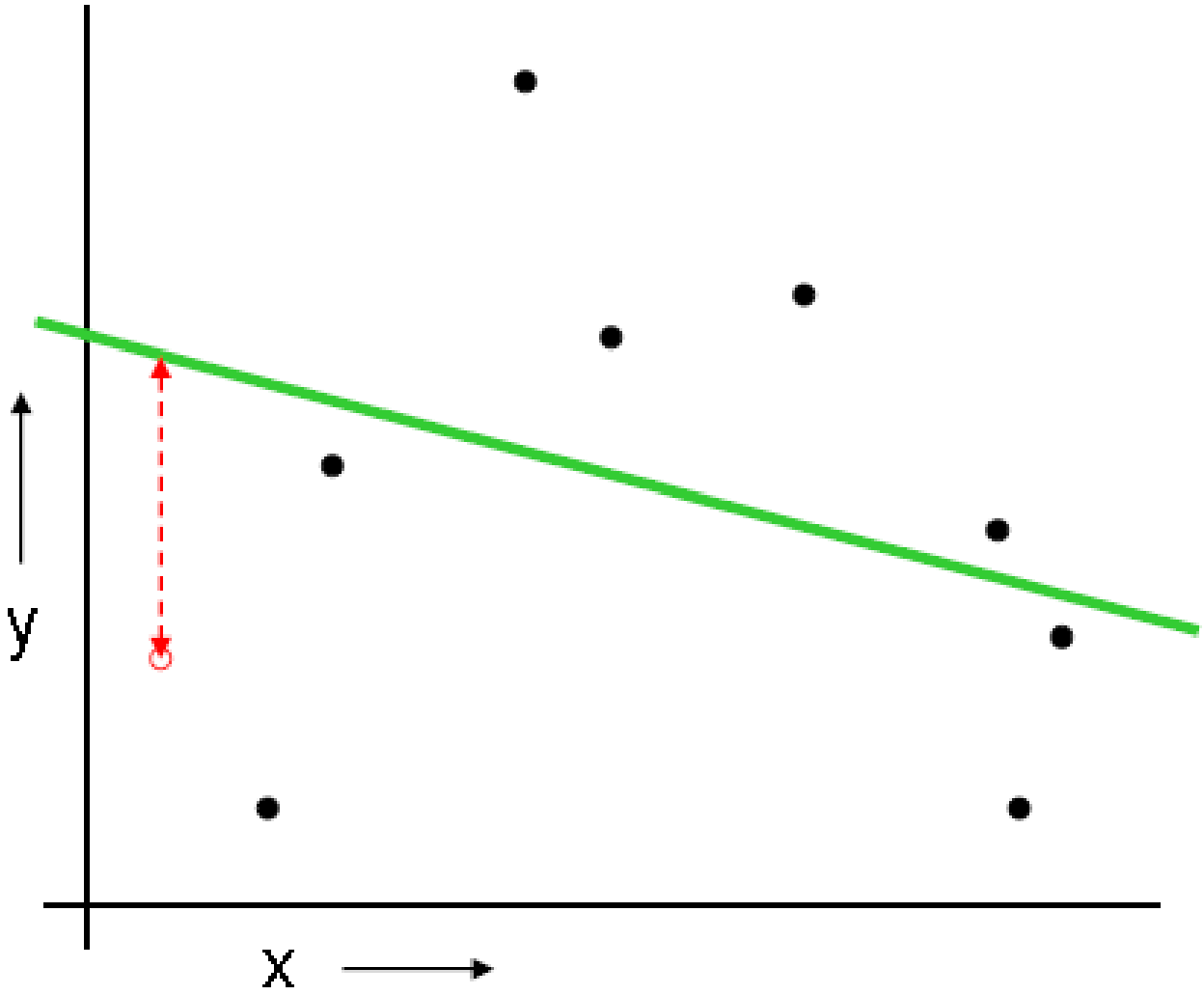
Временно исключаем ее из рассмотрения



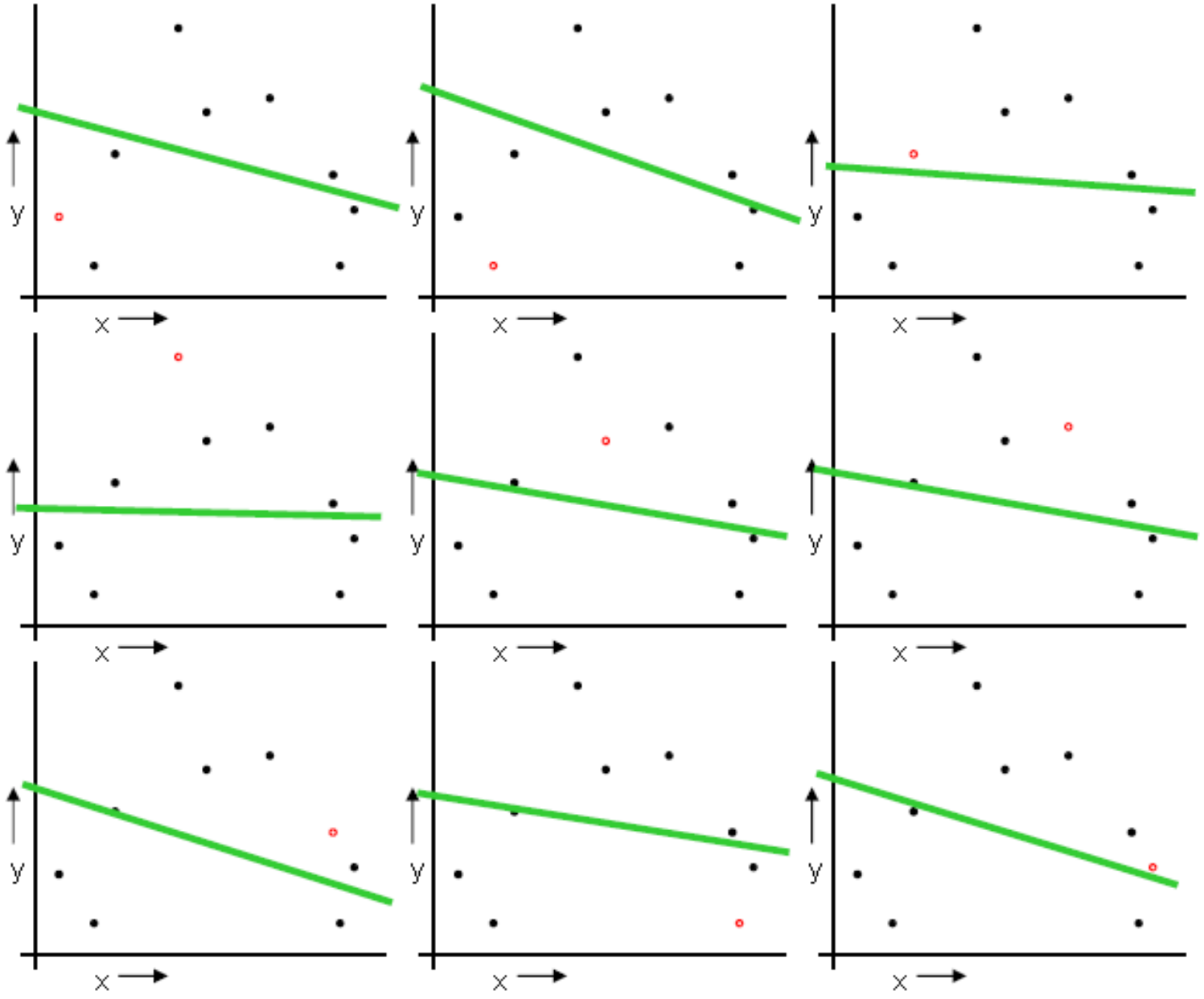
По оставшимся точкам подбираем модель



Определяем ошибку на тестовой выборке



Таким образом перебираем все точки

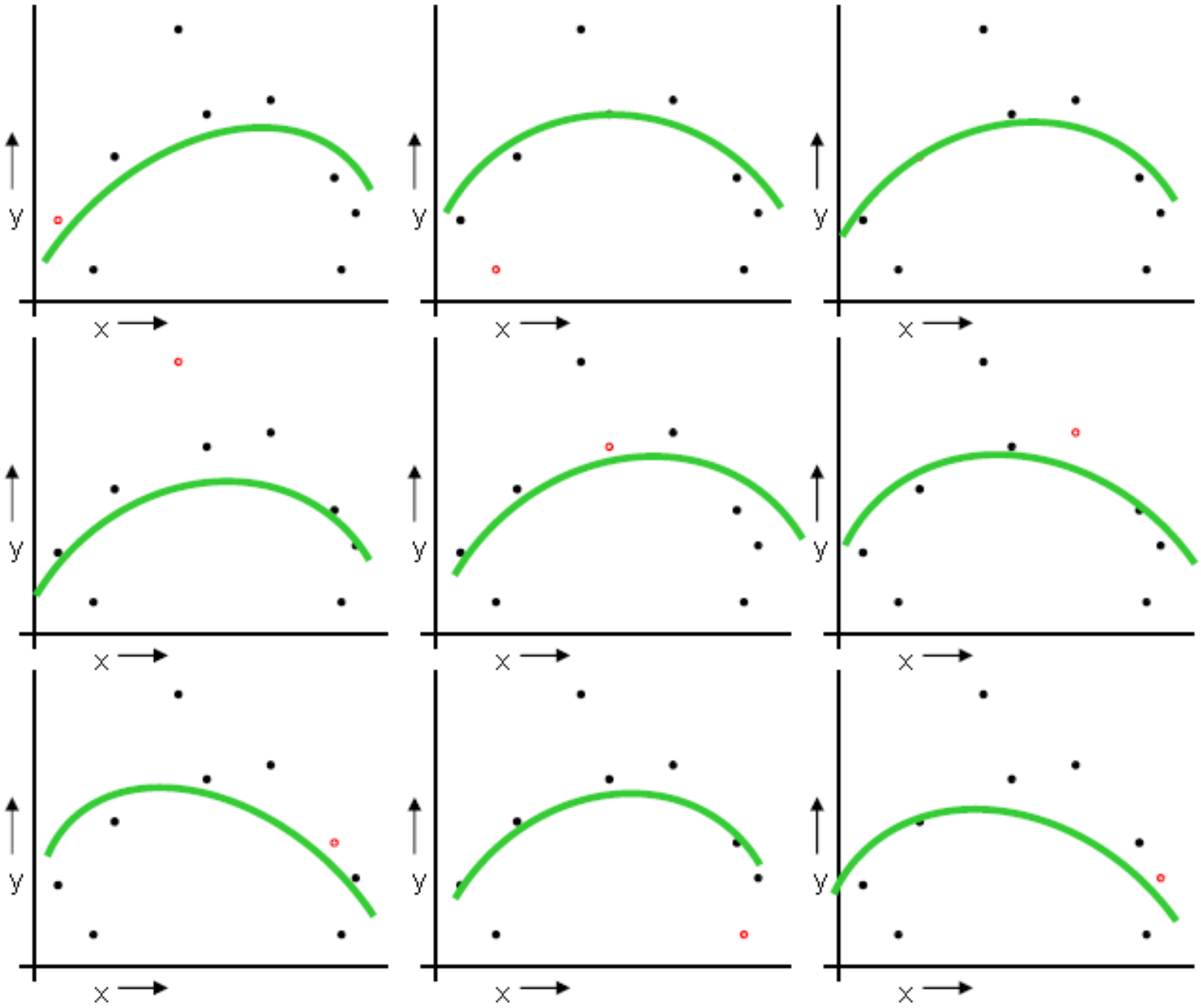


Результат для линейной регрессии

- **Среднее значение квадратов ошибок оказалось равно 2.2**

- **Посчитаем среднее значение квадратов ошибок методом валидации посредством исключенных наблюдений.**

- **Проверим квадратичную регрессионную модель**

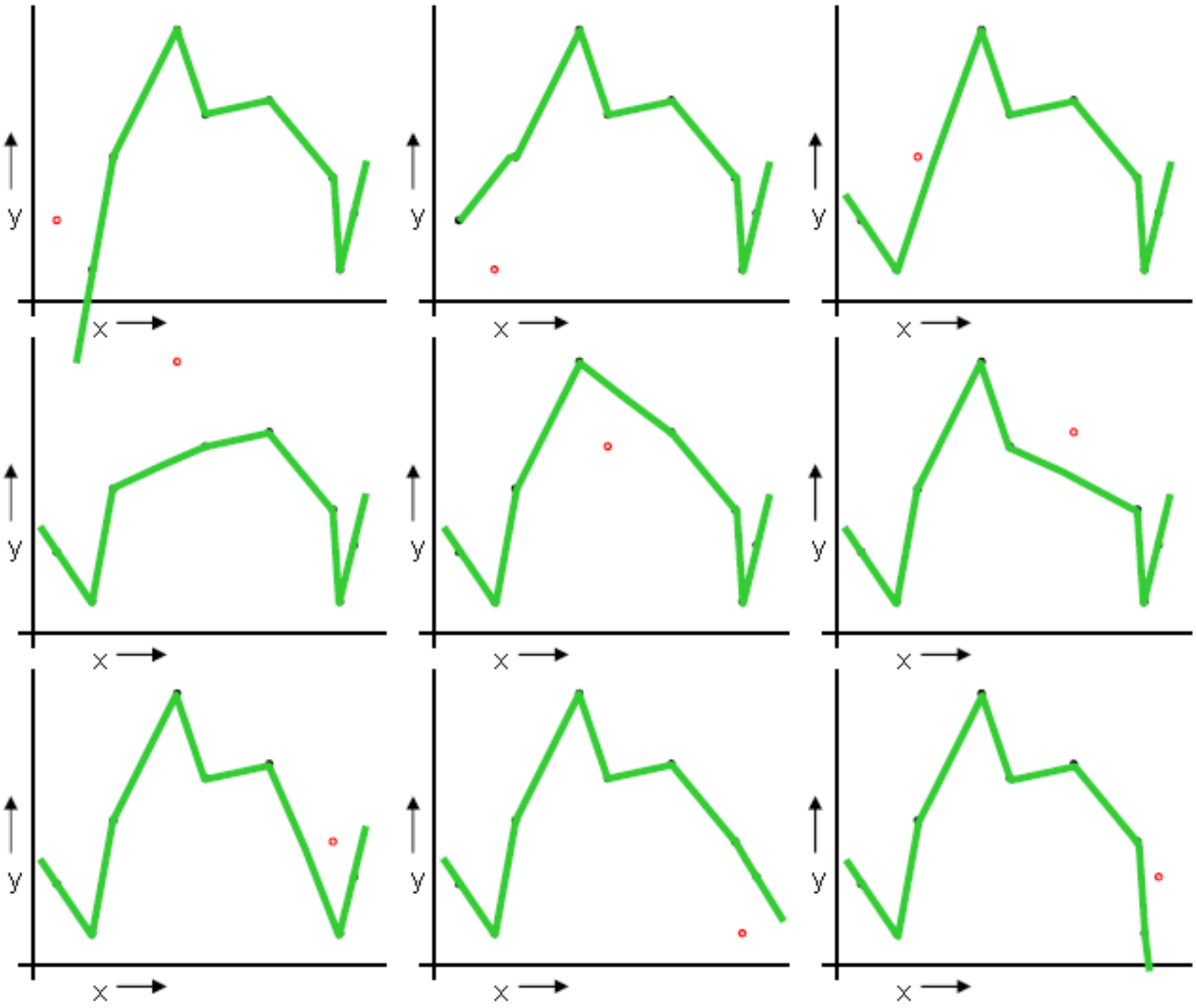

y
x
y
x
y
x
y
x
y
x
y
x
y
x
y
x
y
x

Результат для квадратичной регрессии

- **Среднее значение квадратов ошибок оказалось равно 0.962**

- **Посчитаем среднее значение квадратов ошибок методом валидации посредством исключенных наблюдений.**

- **Проверим модель линейный сплайн.**


y
x
y
x
y
x
y
x
y
x
y
x
y
x
y
x
y
x

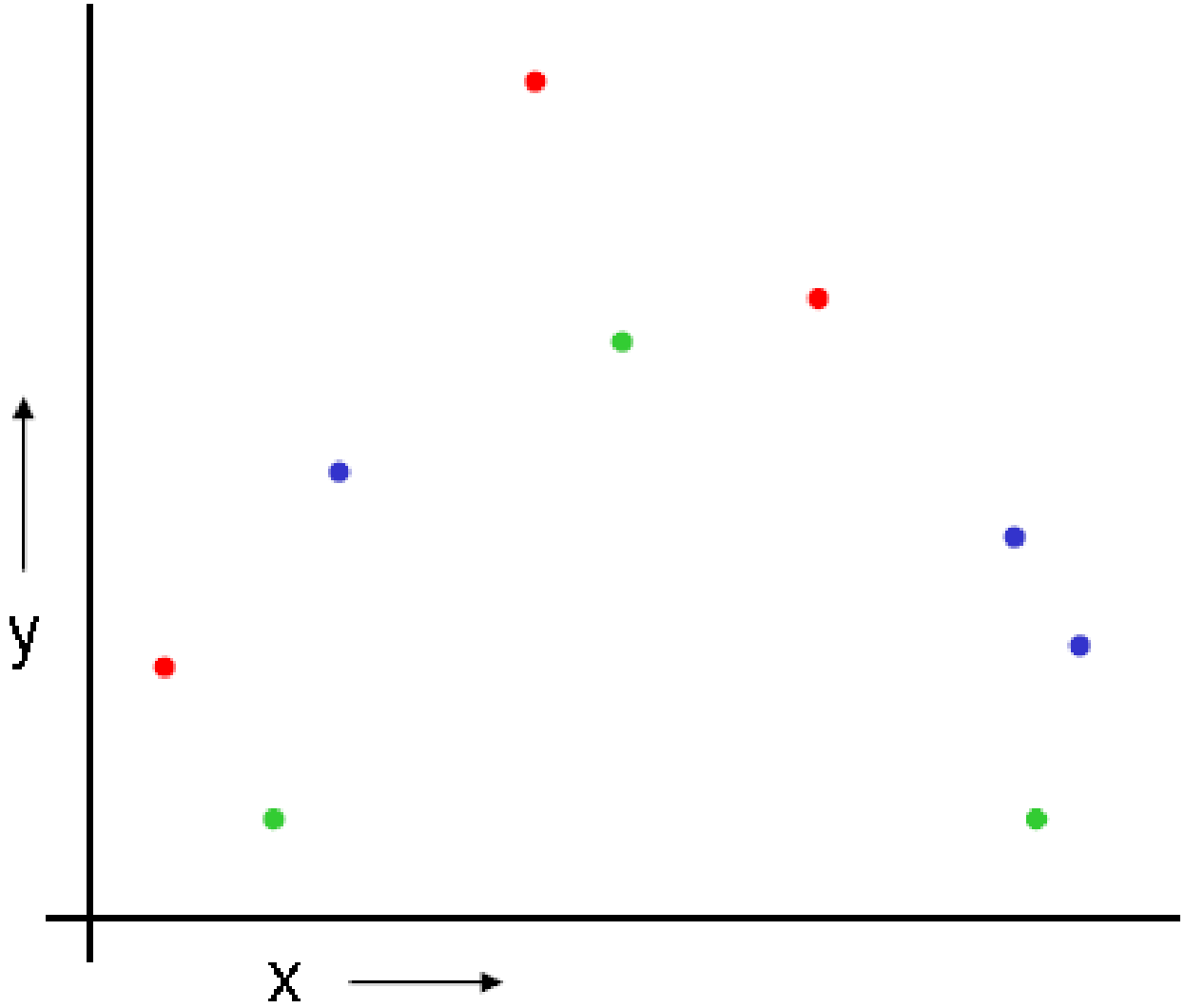
Результат для линейного сплайна

- **Среднее значение квадратов ошибок оказалось равно 3.3**

Метод к-кратной валидации

- **Случайным образом разобьем выборку на к одинаковых частей.**
- **В рассматриваемом примере к=3, точки из разных частей будем отмечать красным, синим и зеленым цветом.**

Получим следующее разбиение

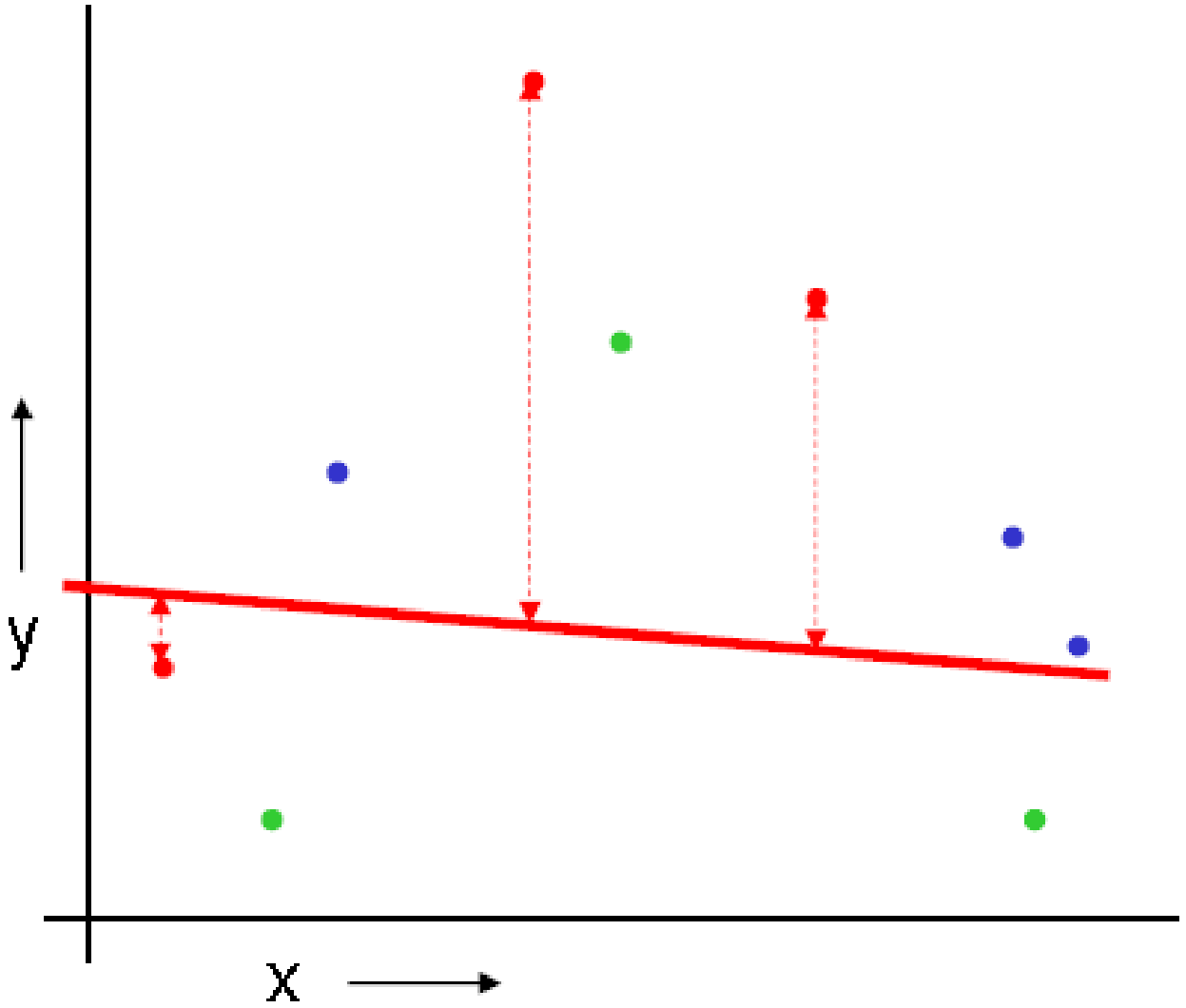


- **Начнем опять с линейной регрессии**

Валидация проводится в к этапов.

- **Первый этап.**

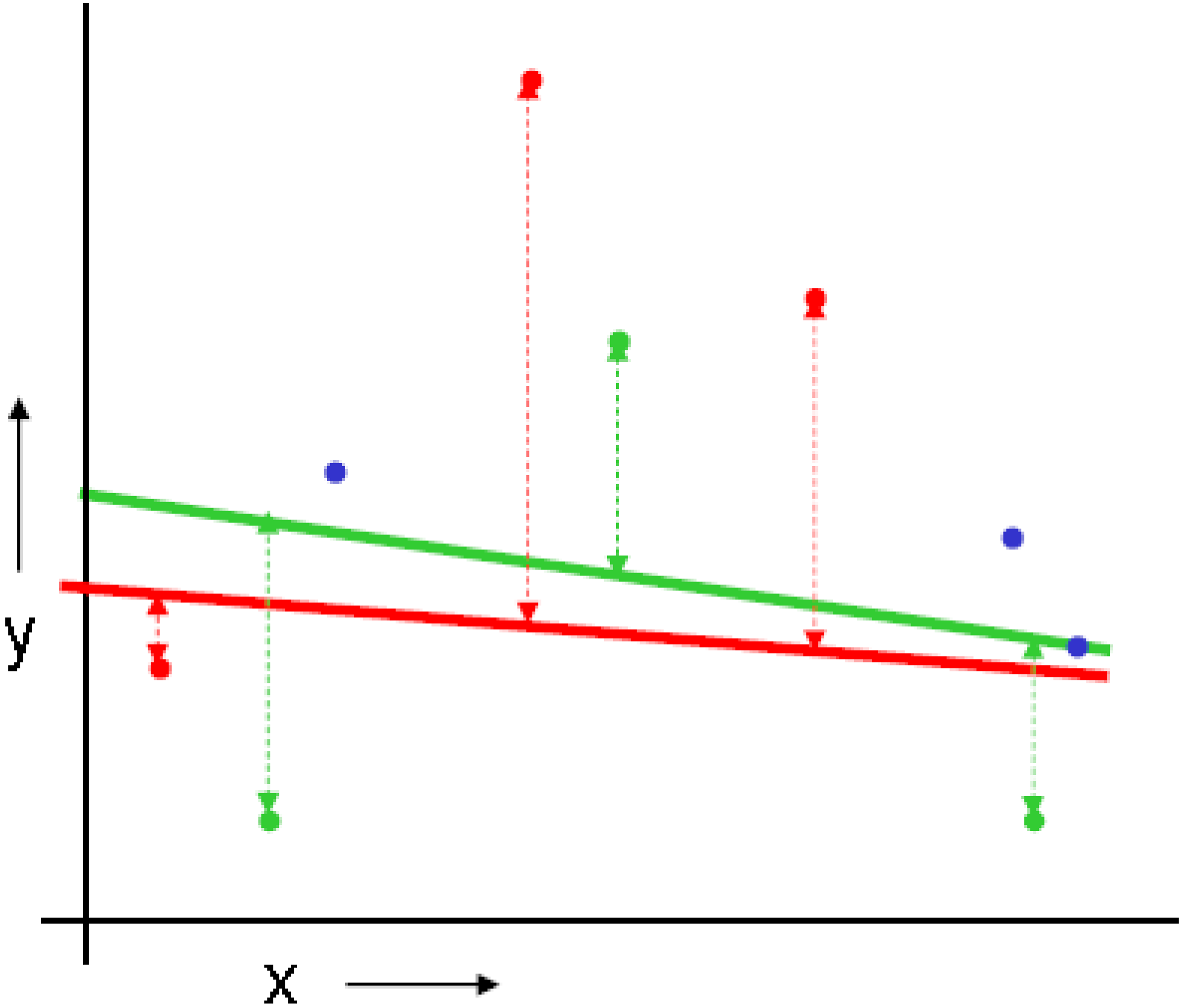
- **Первая часть – тестовая выборка.**
- **Все остальные наблюдения – обучающая выборка.**


y
x

- **Считаем сумму квадратов ошибок для точек из тестового множества.**
- **Определяем обучающую выборку и тестовую выборку заново.**

Второй этап.

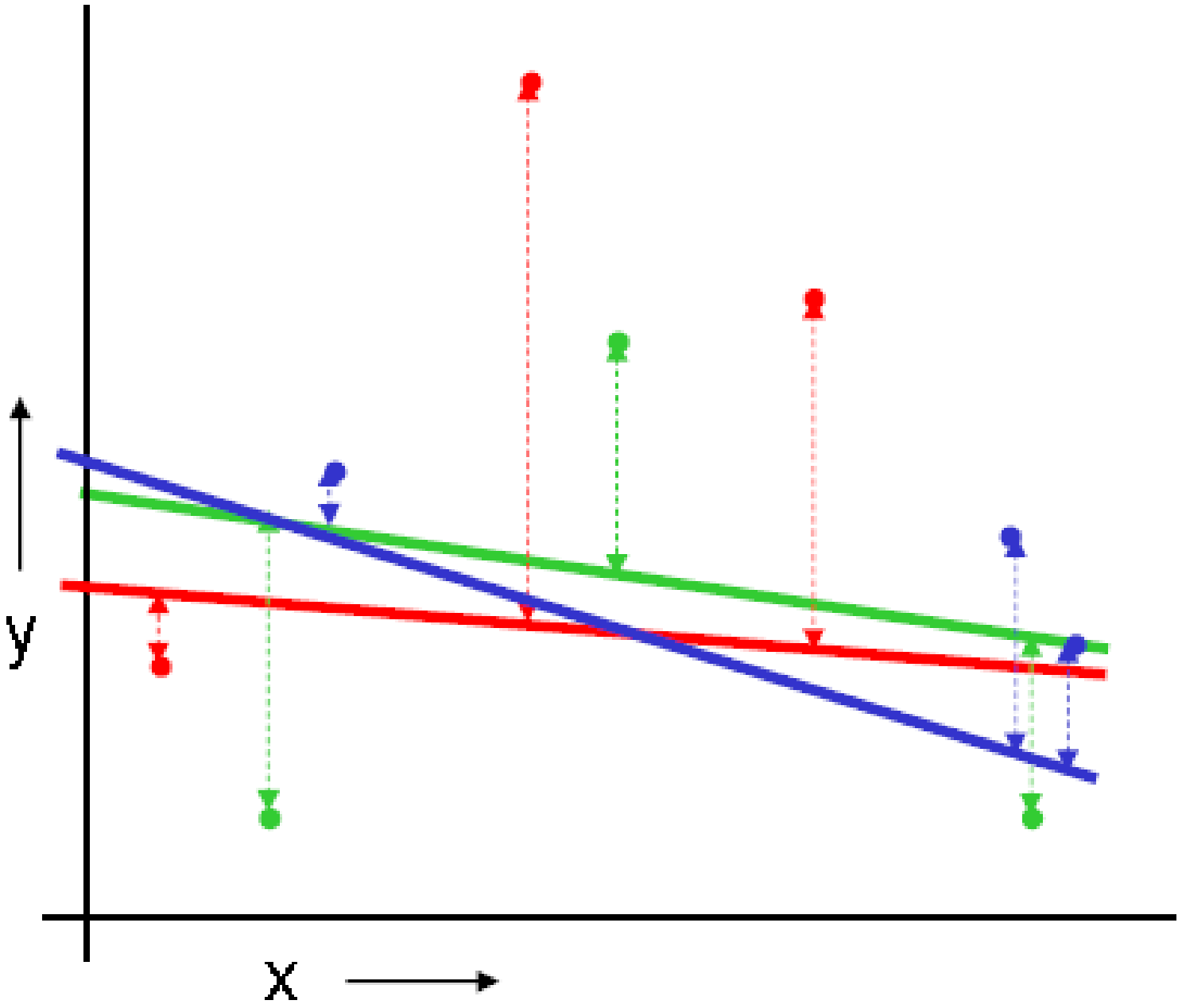
- **Вторая часть – тестовая выборка.**
- **Все остальные наблюдения – обучающая выборка.**


y
x

- **Считаем сумму квадратов ошибок для точек из тестового множества.**
- **Определяем обучающую выборку и тестовую выборку заново.**

Третий этап.

- **Третья часть – тестовая выборка.**
- **Все остальные наблюдения – обучающая выборка.**


y
x

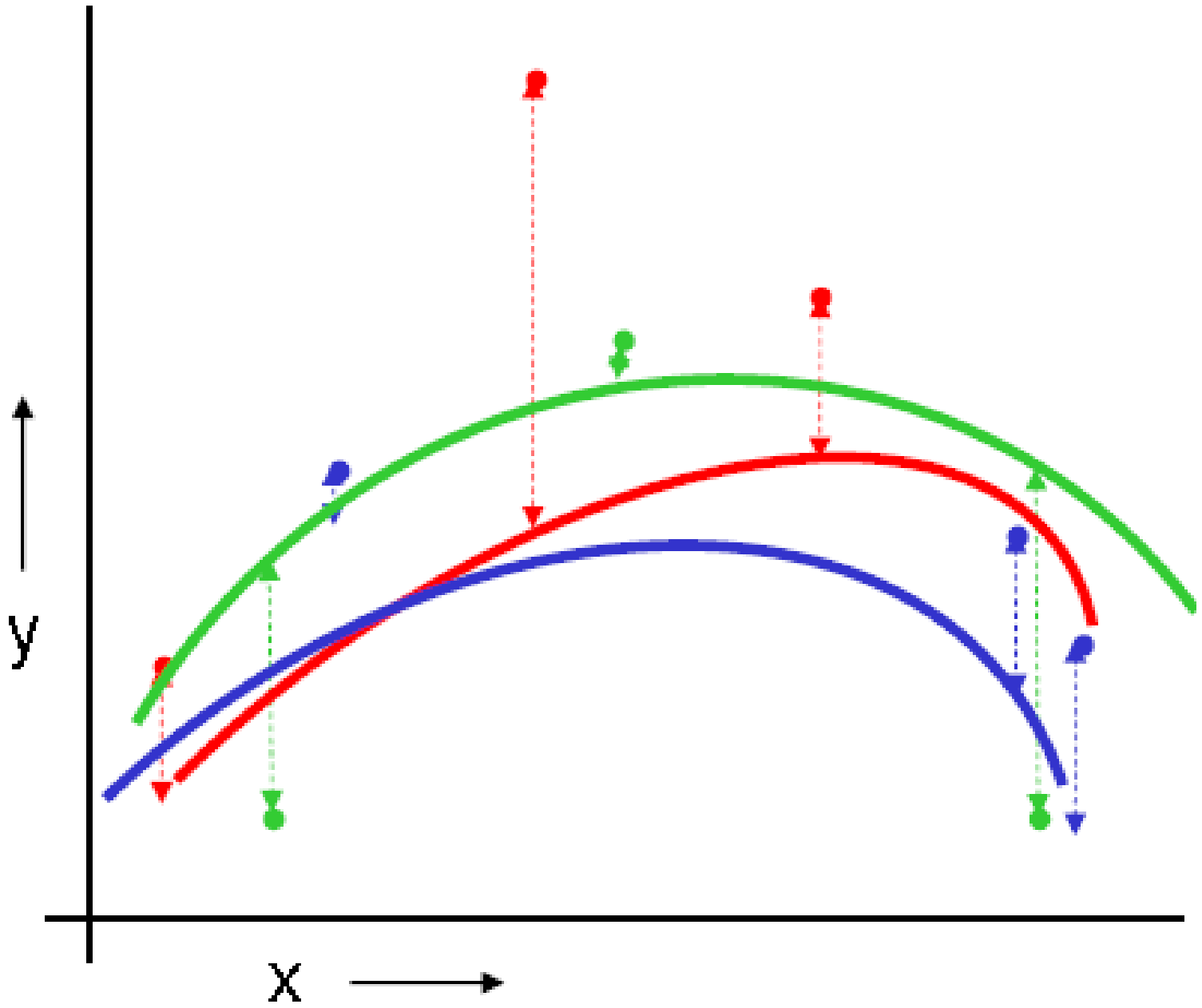
- **Осталось сосчитать среднее значение квадратов ошибок.**
- **Оно оказалось равно 2.05**

- **В примере было к=3**

- **Если в другой задаче значение к будет больше, продолжаем дальше, до к-ой тестовой выборки**

- **Проведем валидацию для квадратичной регрессионной модели**
- **См рисунок на следующем слайде.**

- **среднее значение квадратов ошибок оказалось равно 1.11**


y
x

- **Проведем валидацию для линейного сплайна.**
- **См рисунок на следующем слайде.**

- **среднее значение квадратов ошибок оказалось равно 2.93**

Где точка баланса?