Вариант 27

Упражнение 1

Трапезникова Анастасия

19 Февраль 2021

# Данные варианта.

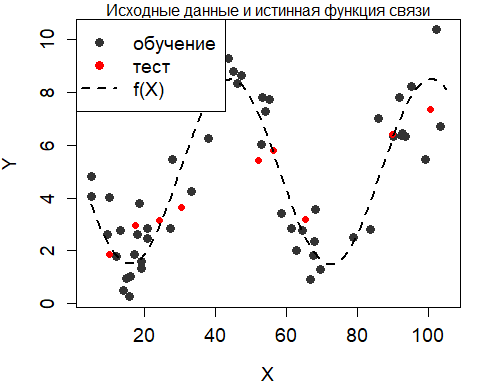
Функция для задачи 1:

Характеристики для задачи 2:

# Задача 1.

На данных своего варианта повторить три графика из первой практики, выбрав число степеней свободы как компромисс между точностью (оценкой ошибки на тестовой выборке) и простотой модели (числом степеней свободы). Все рисунки сохранить в графические файлы в формате png.

Сгенерируем X и Y линейной модели и изобразим исходные данные на графике.



В качестве модели используем сплайны со степенями свободы от 2 (прямая) до 40 (количество узлов равно 2/3 наблюдений). Строим модели с различным количеством степеней свободы и в каждом случае считаем среднеквадратическую ошибку модели на обучающей и тестовой выборках.

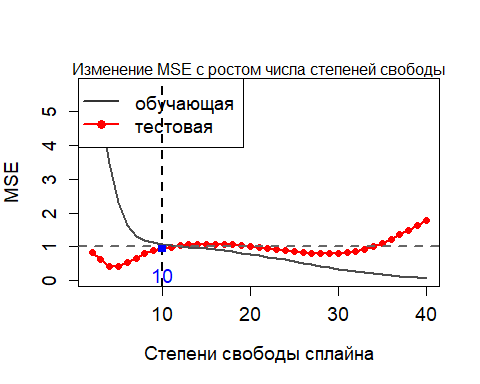
Получаем таблицу вида:

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 5.740766 | 0.8305786 |
| 3 | 4.941521 | 0.6382170 |
| 4 | 3.467131 | 0.4305540 |
| 5 | 2.307132 | 0.4252925 |
| 6 | 1.633279 | 0.5307078 |
| 7 | 1.322383 | 0.6721496 |
| 8 | 1.189844 | 0.7952653 |
| 9 | 1.125319 | 0.8853094 |
| 10 | 1.084036 | 0.9497316 |
| 11 | 1.051278 | 0.9972864 |

Расчёт ошибки на обучающей выборке:

Расчёт ошибки на тестовой выборке:

Изобразим на графике поведение ошибок при различном количестве степеней свободы.

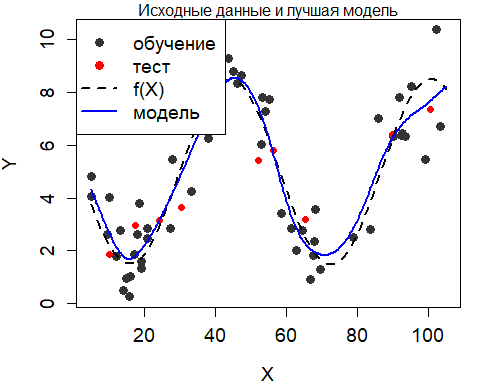


На этом графике:

* При движении слева направо MSE на обучающей выборке (серая кривая) сокращается, потому что с ростом числа степеней свободы растёт число узлов, по которым строится сплайн. При этом модельная кривая подгоняется по всё возрастающему количеству точек и становится всё более гибкой. В результате индивидуальные расстояния от фактических наблюдений за Y до их модельных оценок сокращаются, что приводит к сокращению MSE.
* При движении слева направо MSE на тестовой выборке (красная кривая) сначала медленно сокращается, затем растёт. Нам известна истинная форма связи Y с X, она описывается тригонометрической функцией. Число степеней свободы такой модели равно числу оцениваемых параметров, т.е. 2 (коэффициент перед cos и константа).

Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 11 и равно 1,051278. Визуально по графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 10. Ошибка здесь меньше 1, что ненамного отличается от минимума. Именно df = 10 было выбрано в качестве компромисса между точностью (минимальной MSE на тестовой выборке) и простотой модели (чем меньше степеней свободы, тем модель проще).

График с моделью, выбранной в качестве лучшей, показан на рисунке ниже.

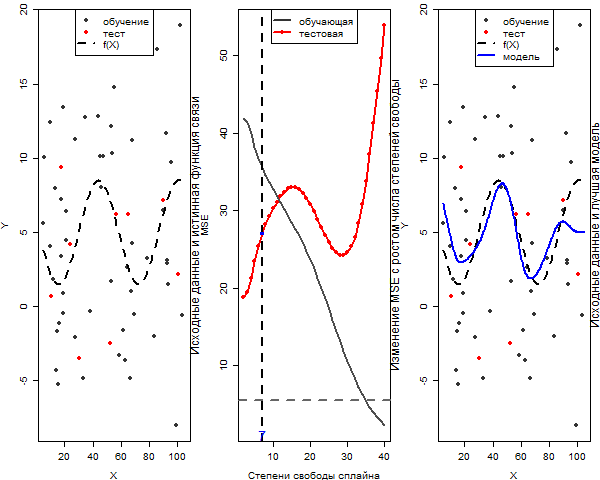


# Задача 2.

*Задание: Решить задачу 1, изменив характеристики данных (стандартное отклонение случайного шума от 2 до 4). Почему при таком изменении данных MSE меняется именно так? Все рисунки сохранить в графические файлы в формате png.*

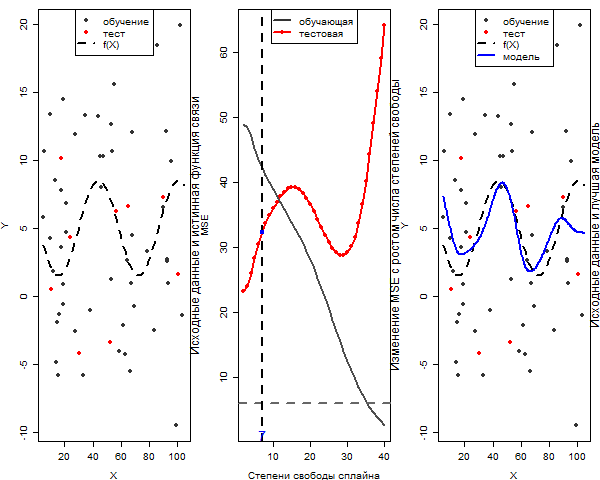
|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 41.83246 | 18.71465 |
| 3 | 41.31806 | 19.45210 |
| 4 | 39.92008 | 21.26702 |
| 5 | 38.16722 | 23.40945 |
| 6 | 36.66305 | 25.38730 |
| 7 | 35.52583 | 26.99173 |
| 8 | 34.55787 | 28.25115 |
| 9 | 33.63672 | 29.29060 |
| 10 | 32.73073 | 30.21463 |
| 11 | 31.84328 | 31.05155 |

## png   
## 2



|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 48.98369 | 23.26735 |
| 3 | 48.47950 | 24.04000 |
| 4 | 47.05080 | 25.97668 |
| 5 | 45.18989 | 28.31238 |
| 6 | 43.54332 | 30.50261 |
| 7 | 42.25420 | 32.29697 |
| 8 | 41.12419 | 33.72240 |
| 9 | 40.03379 | 34.91655 |
| 10 | 38.95648 | 35.99196 |
| 11 | 37.89980 | 36.97420 |

## png   
## 2



|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 56.76313 | 28.33317 |
| 3 | 56.26485 | 29.12738 |
| 4 | 54.79758 | 31.16524 |
| 5 | 52.82021 | 33.68515 |
| 6 | 51.02120 | 36.08861 |
| 7 | 49.56818 | 38.07838 |
| 8 | 48.26227 | 39.67777 |
| 9 | 46.98781 | 41.03646 |
| 10 | 45.72412 | 42.27442 |
| 11 | 44.48335 | 43.41351 |

## png   
## 2

