Упражнение 1

Трапезникова Анастасия

03 Март 2021

# Данные варианта.

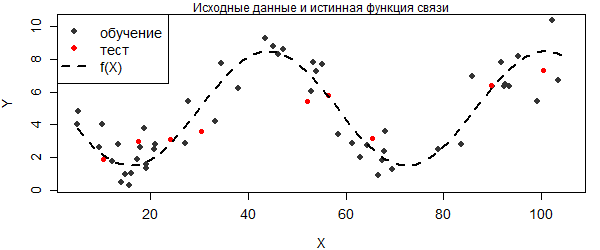
Функция для задачи 1:

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.

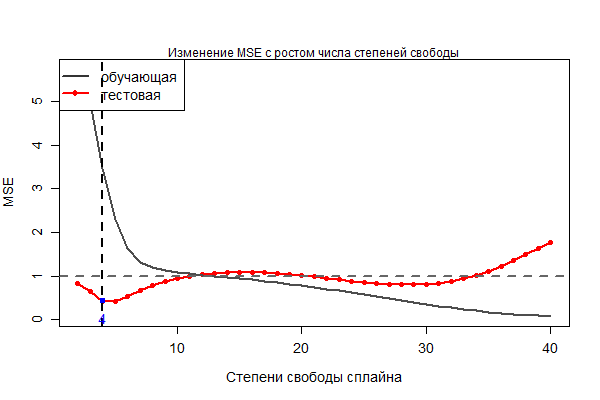
Получаем таблицу вида:

|  |  |  |
| --- | --- | --- |
| 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.

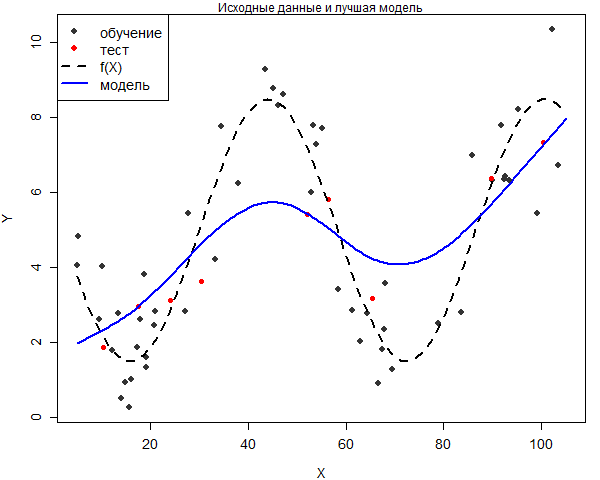


На этом графике:

* 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 синусоидой. Число степеней свободы такой модели может иметь большое число степеней свободы. MSE на тестовой выборке производит слабые колебания вверх-вниз и после 35 степени свободы начинает расти активнее. Этот рост объясняется эффектом переобучения модели: она всё лучше описывает обучающую выборку, и при этом постепенно становится неприменимой ни к одному другому набору наблюдений.

Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 5 и равно 0,425. Визуально по графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 4. Ошибка здесь меньше 1, что ненамного отличается от минимума. Именно df = 4 было выбрано в качестве компромисса между точностью (минимальной MSE на тестовой выборке) и простотой модели (чем меньше степеней свободы, тем модель проще). При выборе df = 11 (лучший вид модели, наиболее похожий на f(x)) ошибка составляет ~ 20% от среднего Y, а при 4 она не превышает 8%, так что для компромисса мы пожертвуем видом синусоиды.

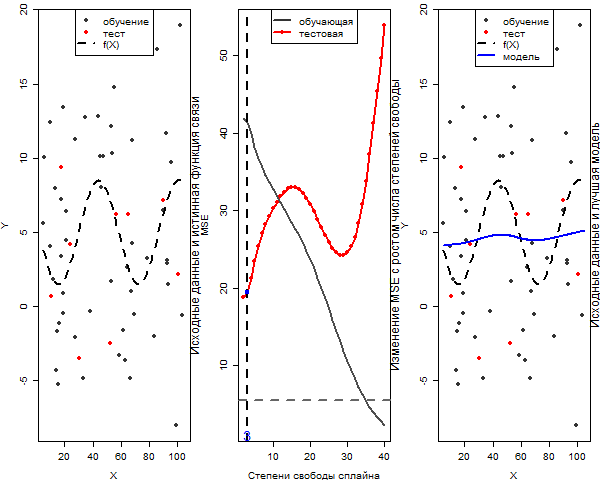
График с моделью, выбранной в качестве лучшей, показан на рисунке ниже.



# Задача 2.

*Задание: Решить задачу 1, изменив характеристики данных (стандартное отклонение случайного шума от 5,5 до 6,5). Почему при таком изменении данных MSE меняется именно так? Все рисунки сохранить в графические файлы в формате png.*

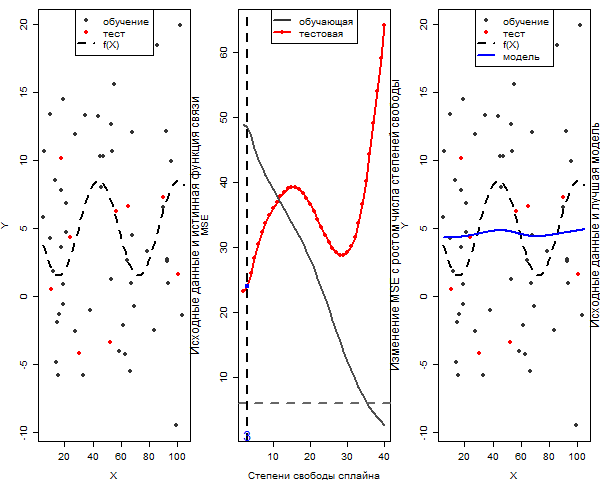
|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 41.83246 | 18.71465 |
| 3 | 41.31806 | 19.45210 |
| 4 | 39.92008 | 21.26702 |
| 5 | 38.16722 | 23.40945 |
| 6 | 36.66305 | 25.38730 |
| 7 | 35.52583 | 26.99173 |
| 8 | 34.55787 | 28.25115 |
| 9 | 33.63672 | 29.29060 |
| 10 | 32.73073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



Неустранимая ошибка модели равняется 5,5, можно обратить внимание, что рассеивание точек стало больше. Ошибка обучающей выборки ниже неустранимой ошибки только при степени свободы больше 35.

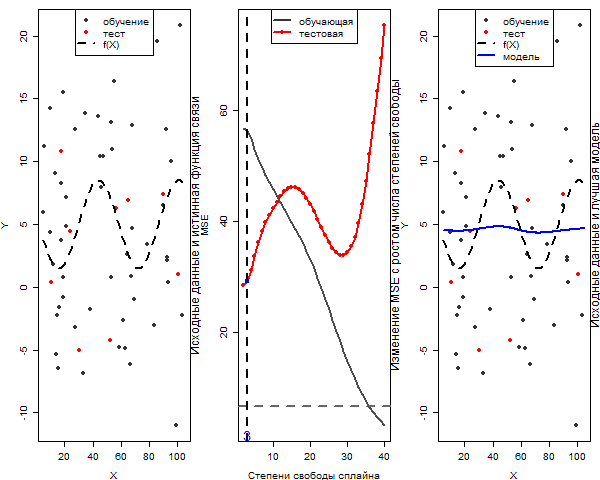
Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 2 и равно 18,71. Визуально по графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 3. Ошибка здесь больше 5,5, что отличается от минимума. Именно df = 3 было выбрано в качестве компромисса между точностью (минимальной MSE на тестовой выборке) и простотой модели (чем меньше степеней свободы, тем модель проще).

|  |  |  |
| --- | --- | --- |
| 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 |



Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 2 и равно 23,27. Визуально по графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 3. Ошибка здесь больше 6.

|  |  |  |
| --- | --- | --- |
| 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 |



Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 2 и равно 28,3. Визуально по графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 3. Ошибка здесь больше 6,5.

**Выводы:**

При такой большой неустранимой ошибке невероятно сложно повторить синусоиду при компромиссе между точностью и простотой модели. Тестовая выборка отдаляется от неё, а обучающая пересекает при увеличении числа степеней свободы.