Упражение № 1

Кориба Марина

01 03 2021

# Данные

Функция для задачи 1:

Характеристики для задачи 2:

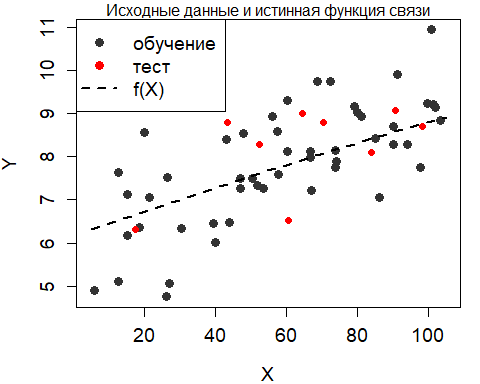
# Задача 1.

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.

#### Генерация данных

###### График 1: Исходные данные на график



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:

## train test   
## 0.85 0.70

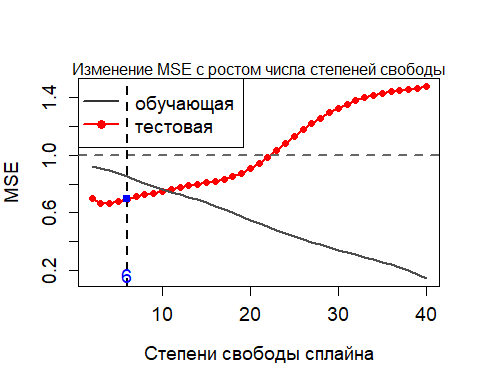
|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 0.9252216 | 0.7012261 |
| 3 | 0.9054864 | 0.6622866 |
| 4 | 0.8918758 | 0.6621592 |
| 5 | 0.8727891 | 0.6772048 |
| 6 | 0.8503551 | 0.6958308 |
| 7 | 0.8262665 | 0.7118193 |
| 8 | 0.8027568 | 0.7244094 |
| 9 | 0.7814953 | 0.7361357 |
| 10 | 0.7624297 | 0.7487297 |
| 11 | 0.7446615 | 0.7620765 |

Расчёт ошибки на обучающей выборке:

Расчёт ошибки на тестовой выборке:

Изобразим на графике поведение ошибок при различном количестве степеней свободы.

###### График 2: Зависимость MSE от гибкости модели



На этом графике:

При движении слева направо MSE на обучающей выборке (серая кривая) постоянно сокращается, потому что с ростом числа степеней свободы расчёт 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.

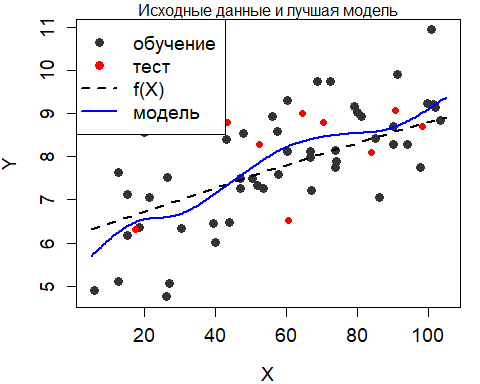
При движении слева направо MSE на тестовой выборке (красная кривая) сначала чуть-чуть сокращается, затем растёт. Нам известна истинная форма связи Y с X, она описывается линейной функцией. Число степеней свободы такой модели равно числу оцениваемых параметров, т.е. 2 (коэффициент перед X и константа). Поэтому резкое падение ошибки на тестовой выборке при небольшом числе степеней свободы связано с тем, что модель приближается по гибкости к истинной функции связи. Затем MSE на тестовой выборке довольно долго остаётся стабильной, а затем начинает расти. 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 3 и равно 0.6622866. Визуально по

графику мы можем установить, что первое значение MSEТЕСТ, близкое к стабильно низким, соответствует df = 5. Ошибка здесь равна 1, что ненамного отличается от минимума. Именно df = 5 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.

###### График 3: Лучшая модель (компромисс между гибкостью и точностью)

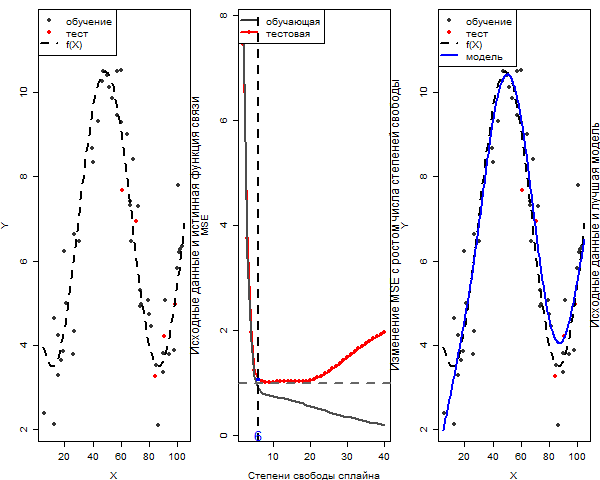


# Задача 2.

*Задание: Решить задачу 1, изменив характеристики данных. Почему при таком изменении данных MSE меняется именно так? Все рисунки сохранить в графические файлы в формате png.*

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 7.8165627 | 7.456438 |
| 3 | 3.3691356 | 3.746622 |
| 4 | 1.7958460 | 1.954609 |
| 5 | 1.1317229 | 1.238297 |
| 6 | 0.8902440 | 1.059573 |
| 7 | 0.8071887 | 1.025888 |
| 8 | 0.7732972 | 1.018863 |
| 9 | 0.7537739 | 1.016921 |
| 10 | 0.7386679 | 1.017792 |
| 11 | 0.7245951 | 1.020413 |

## png   
## 2

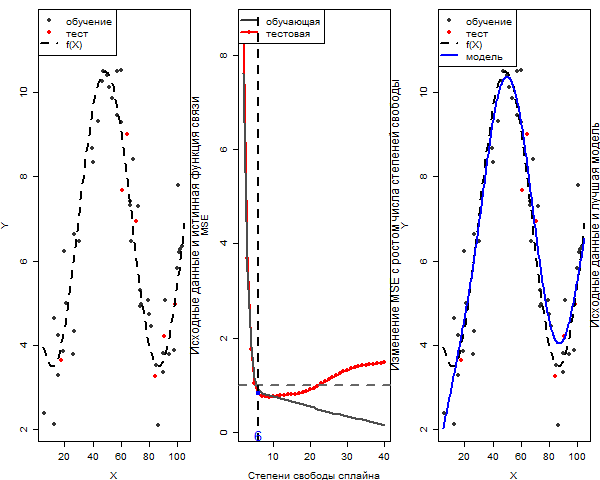


Наименьшее значение MSE на тестовой выборке соответствует числу степеней свободы 9 и равно 1.01692. По графику мы можем установить, что первое значение MSEТЕСТ df = 8.

df = 8 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 | 7.6096839 | 8.6304752 |
| 3 | 3.4224596 | 3.6954224 |
| 4 | 1.8377315 | 1.7695919 |
| 5 | 1.1660393 | 1.0531505 |
| 6 | 0.9225087 | 0.8367075 |
| 7 | 0.8375594 | 0.7762173 |
| 8 | 0.8012519 | 0.7591441 |
| 9 | 0.7790661 | 0.7568933 |
| 10 | 0.7610404 | 0.7618866 |
| 11 | 0.7440446 | 0.7706571 |

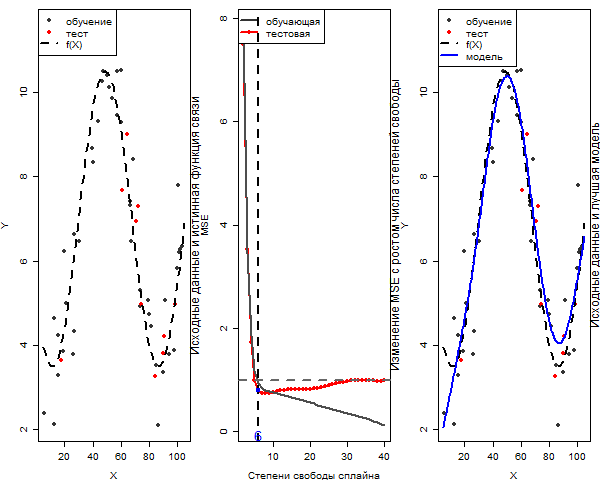
## png   
## 2



Наименьшее значение MSE на тестовой выборке df = 8 и равно 0.7591441. По графику можно заметить, что первое значение MSEТЕСТ df = 7.

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 7.8600057 | 7.4963304 |
| 3 | 3.4445124 | 3.5240588 |
| 4 | 1.8651109 | 1.7201305 |
| 5 | 1.1804148 | 1.0001780 |
| 6 | 0.9311171 | 0.7956994 |
| 7 | 0.8449055 | 0.7476775 |
| 8 | 0.8070429 | 0.7406508 |
| 9 | 0.7831948 | 0.7470332 |
| 10 | 0.7636325 | 0.7594773 |
| 11 | 0.7454694 | 0.7743245 |

## png   
## 2



Наименьшее значение MSE на тестовой выборке df = 8 и равно 0.740650. По графику можно предположить, что значение MSEТЕСТ, соответствует df = 6.

* Почему при таком изменении данных MSE меняется именно так?

в нашем задании изначально имеется доволно большой процент выборки (0,85). Изменения данных происходят имеено таким образом, так как мы меняем долю выборки (доля обучающей выборки меняется в иапозоне от 0.9 до 0.8), так же имеет значение тот факт, что модель пытается построить функцию по всем имеющимся точкам.