Вариант 26

Степанов Герман

24 Февраль 2021

output: word\_document

# Данные варианта.

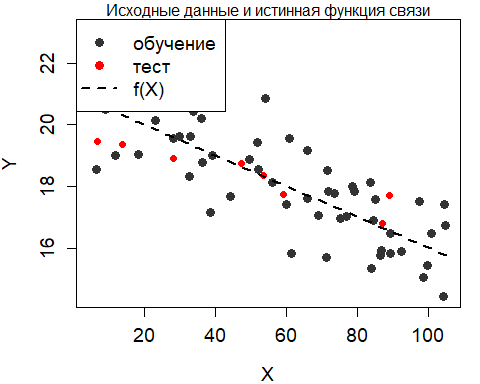
Функция для задачи 1:

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.



## train test   
## 1.02 32.15

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 | 1.105949 | 1.095611 |
| 3 | 1.077058 | 1.226485 |
| 4 | 1.060668 | 1.358124 |
| 5 | 1.044323 | 1.469753 |
| 6 | 1.023871 | 1.590716 |
| 7 | 1.000464 | 1.726380 |

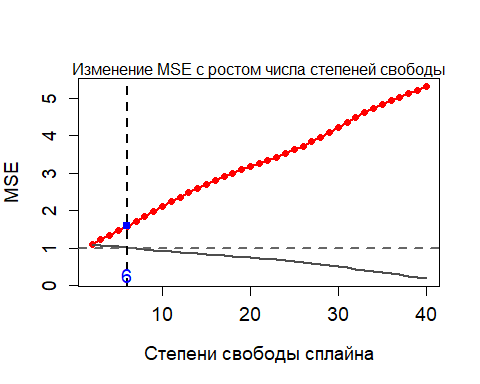
Наименьшая MSE на тестовой выборке равна 1.1 и достигается при df = 2.

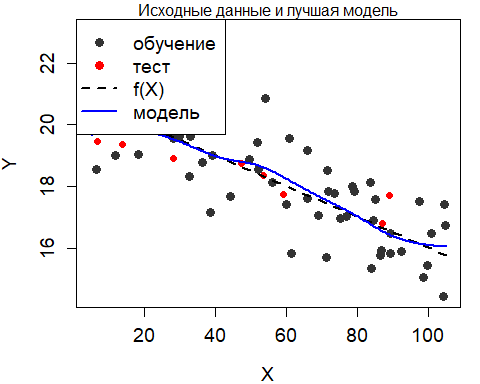
Компромисс между точностью и сложностью модели при df = 6, MSE = 1.59.

Расчёт ошибки на обучающей выборке:

Расчёт ошибки на тестовой выборке:

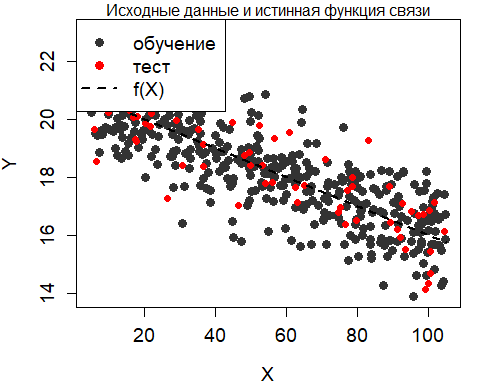
Изобразим на графике поведение ошибок при различном количестве степеней свободы.





# Задача 2.

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

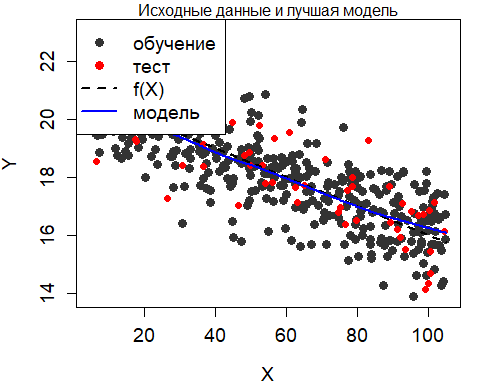
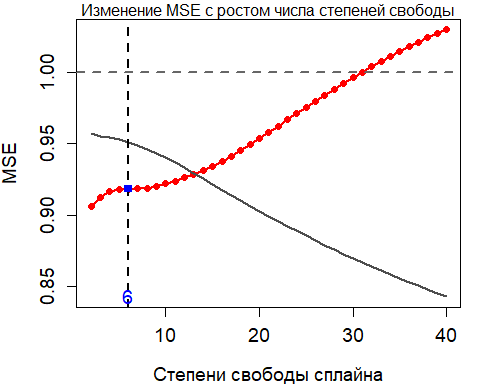


## train test   
## 0.95 29.59

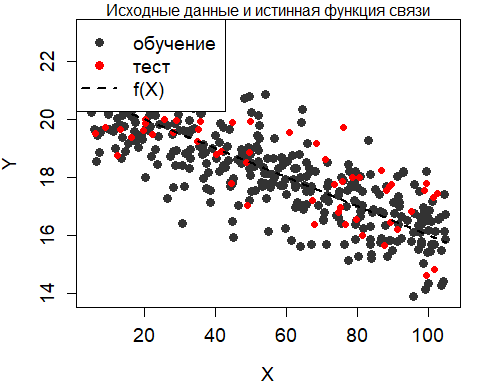
|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 0.9575806 | 0.9057835 |
| 3 | 0.9553854 | 0.9121338 |
| 4 | 0.9542213 | 0.9162893 |
| 5 | 0.9529128 | 0.9179493 |
| 6 | 0.9512078 | 0.9182433 |
| 7 | 0.9490543 | 0.9183545 |

## Наименьшая MSE на тестовой выборке равна 0.91 и достигается при df = 2.

## Компромисс между точностью и сложностью модели при df = 6, MSE = 0.92.



# При *‘n.all’* = 400

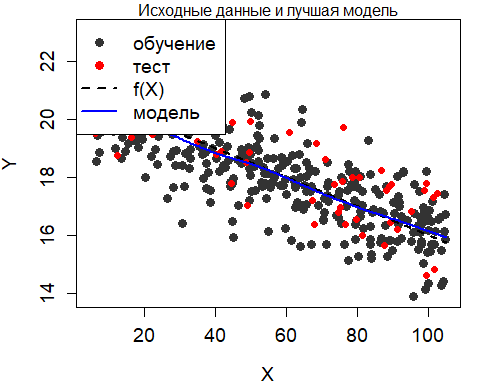
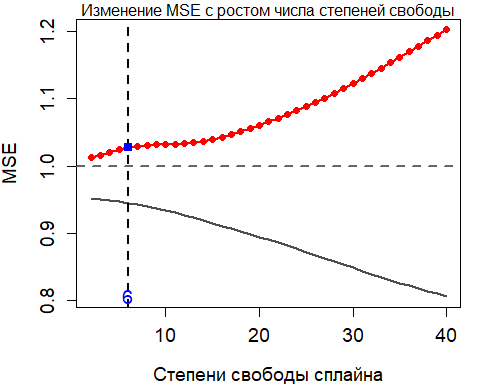


## train test   
## 0.95 24.31

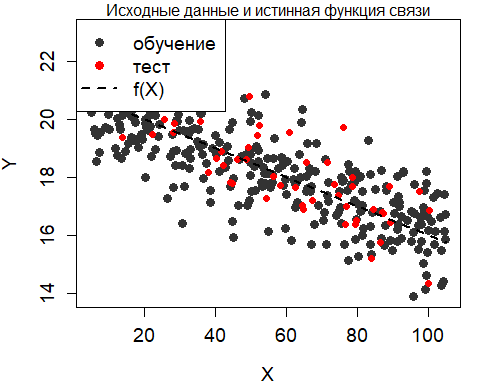
|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 0.9516218 | 1.012655 |
| 3 | 0.9503059 | 1.016096 |
| 4 | 0.9491628 | 1.020133 |
| 5 | 0.9474430 | 1.024537 |
| 6 | 0.9452935 | 1.027922 |
| 7 | 0.9428767 | 1.030029 |

## Наименьшая MSE на тестовой выборке равна 1.01 и достигается при df = 2.

## Компромисс между точностью и сложностью модели при df = 6, MSE = 1.03.



# При *‘n.all’* = 350

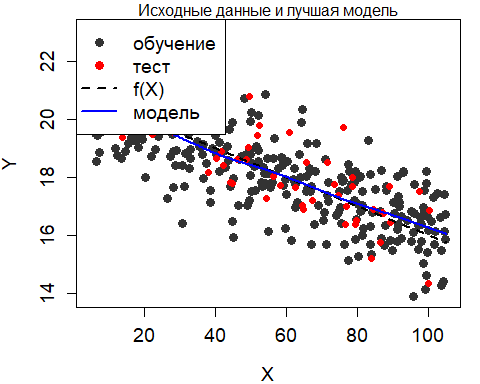
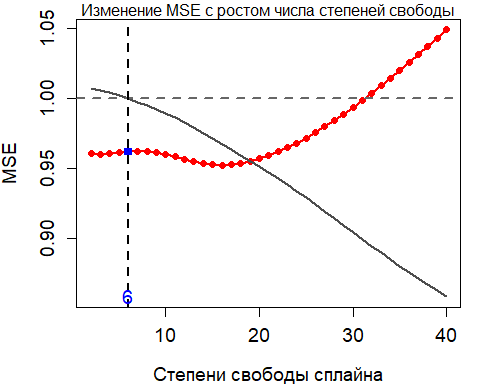


## train test   
## 1.00 26.88

|  |  |  |
| --- | --- | --- |
| df | MSE.train | MSE.test |
| 2 | 1.0066280 | 0.9603936 |
| 3 | 1.0052795 | 0.9594778 |
| 4 | 1.0040884 | 0.9602535 |
| 5 | 1.0020512 | 0.9611364 |
| 6 | 0.9995064 | 0.9616824 |
| 7 | 0.9969892 | 0.9619300 |

## Наименьшая MSE на тестовой выборке равна 0.95 и достигается при df = 16.

## Компромисс между точностью и сложностью модели при df = 6, MSE = 0.96.



**Вывод:**

Среднее квадратическое отклонение (MSE) - мера рассеяния значений случайной величины относиетльно её среднего значения (математического ожидания) - корень квадратный из дисперсии случайной величины (второго центрального момента распределения). При росте количества наблюдений, MSE снижается, что помогает нам сделать более точный выбор.