Предсказание содержания белка по результату спектрограммы

Алексей Усачев, Валерий Бабушкин

Данные

Тренировочная выбрка: 5620 строк (содержание белка) по 330 параметров (показатели спектрометра на разных частотах). Все значения float.

Тестовая выборка: 458 строк, то же количество параметров.

ВАЖНО:объекты тестового множества относятся к другому подклассу, чем объекты обучающего

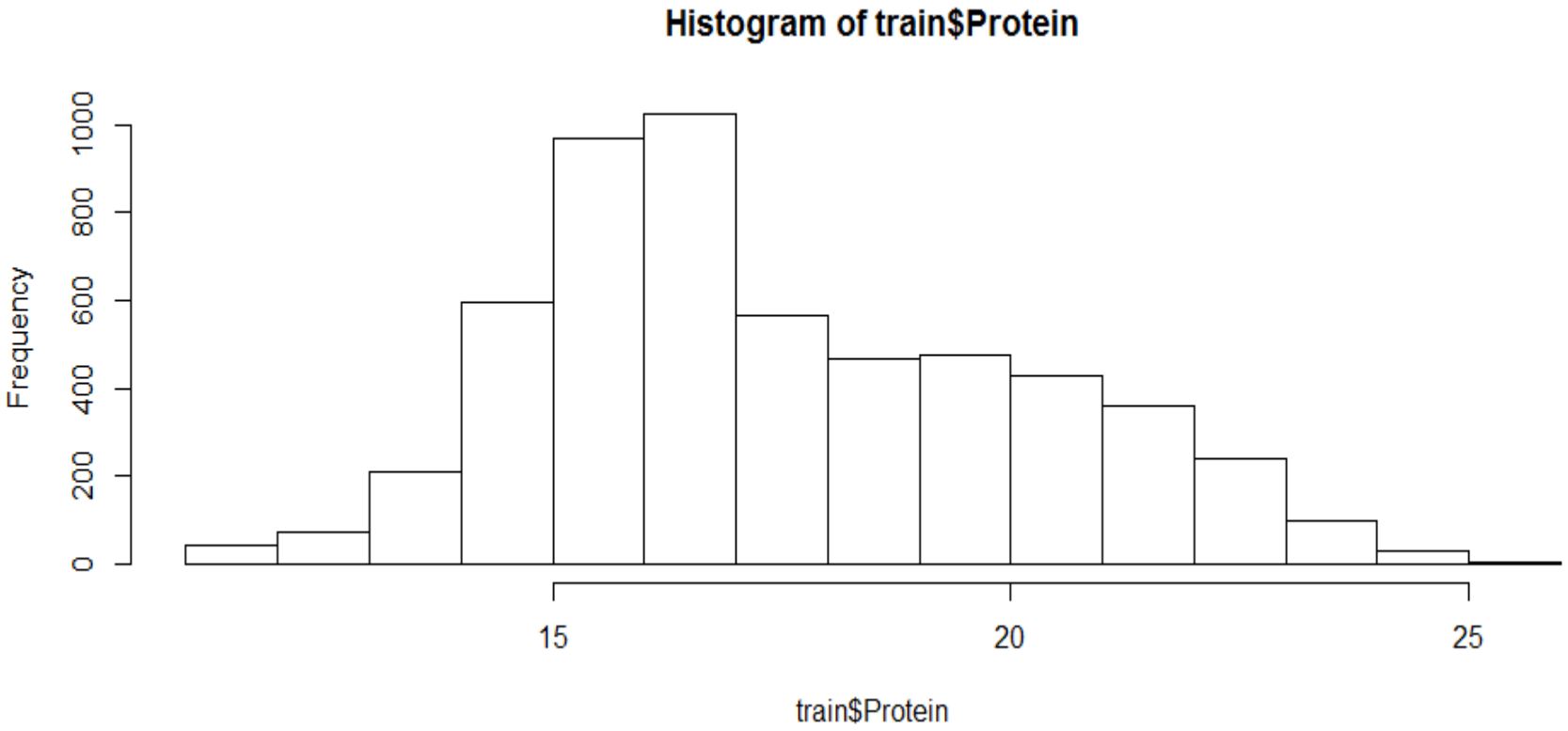
Для очистки данных использовался признак: выбросами считаются значения, отличающиеся от среднего больше, чем СКО*3. После нее осталось 5606 значений.

Для повторной очистки использовалось расстояние Махаланобиса.

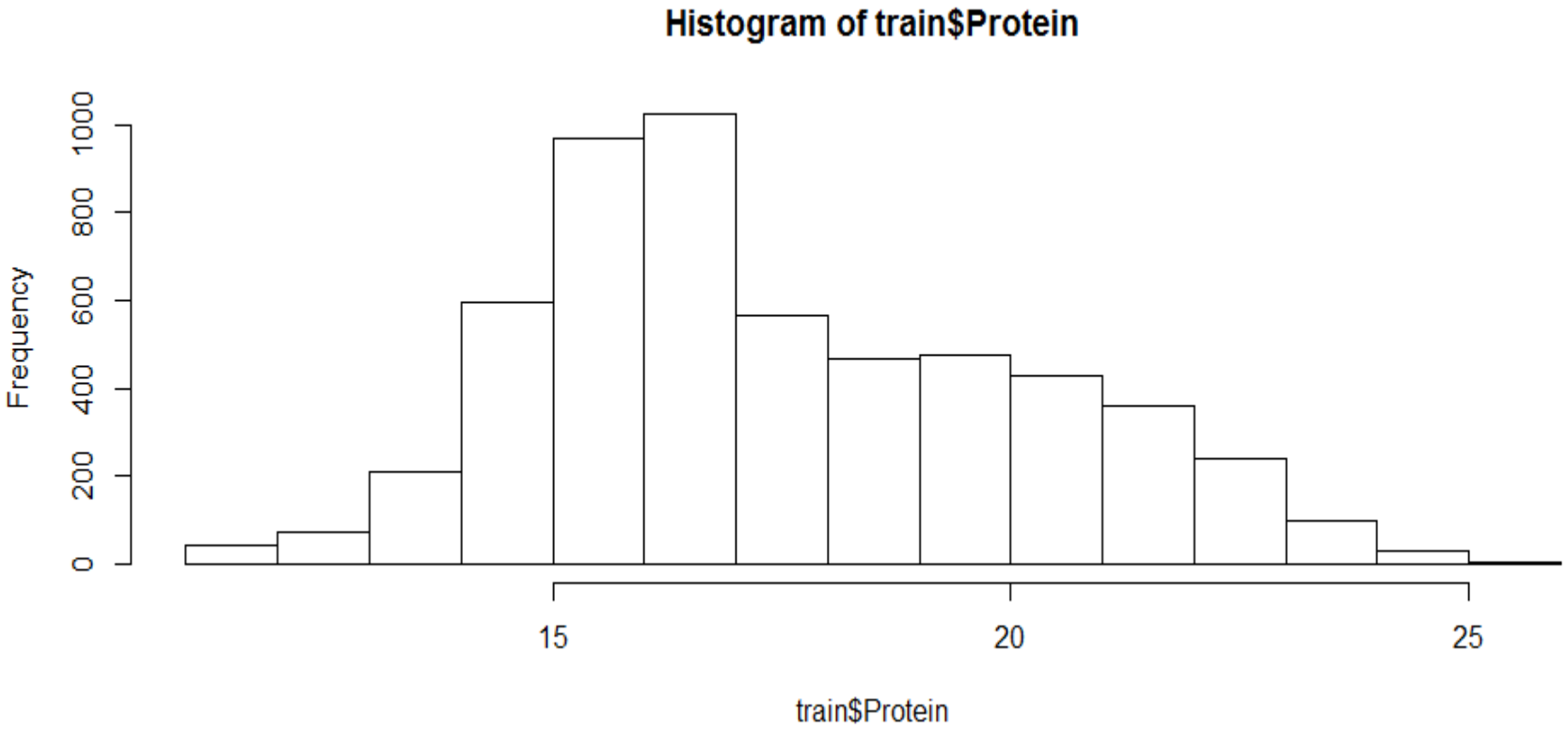
Очистка привела к некоторому улучшению моделию

Распределение образцов в тренировочном наборе

Распределение образцов в тренировочном наборе после очистки



Распределение образцов в тренировочном наборе после очистки



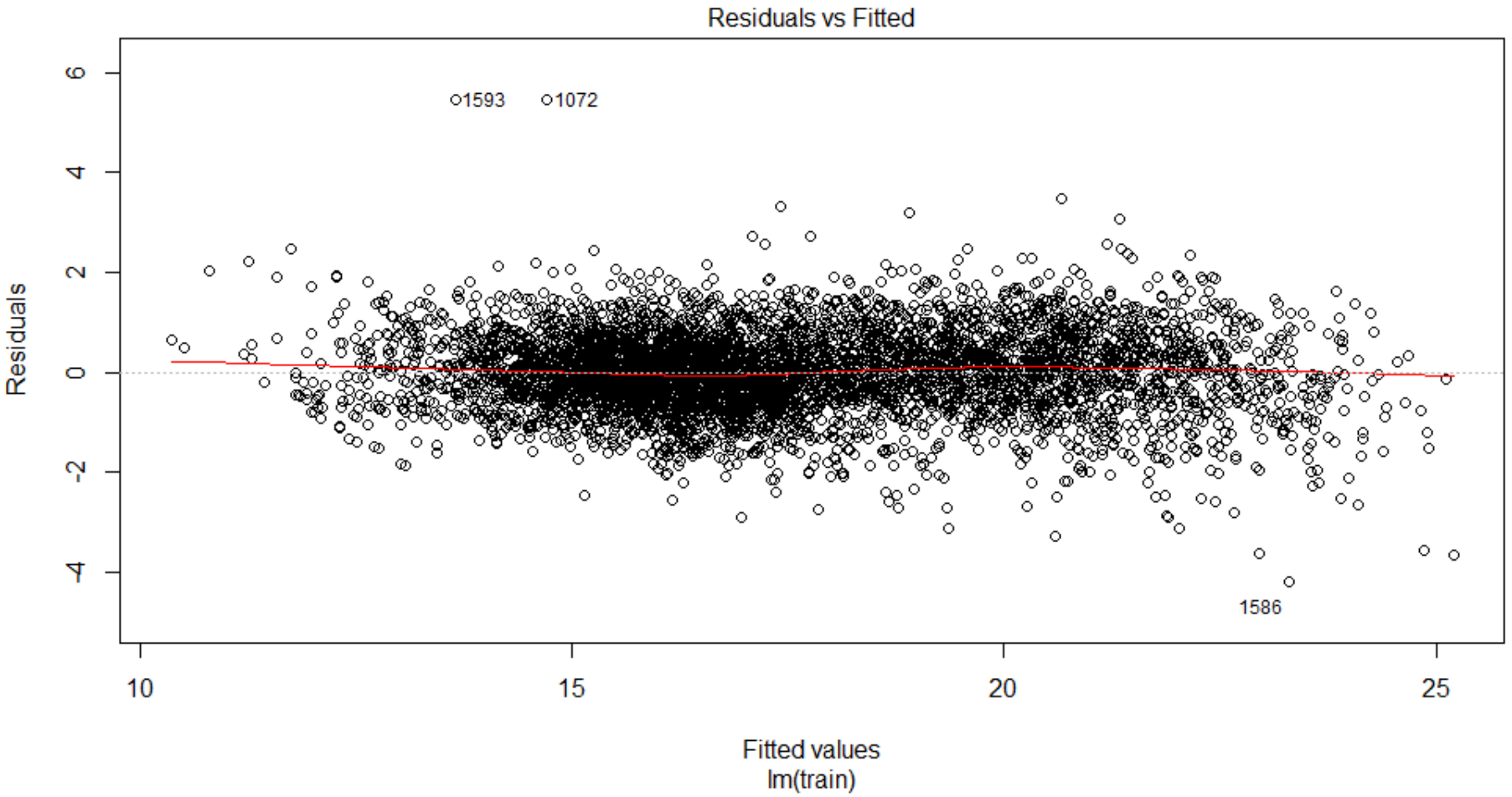
Задача регрессии

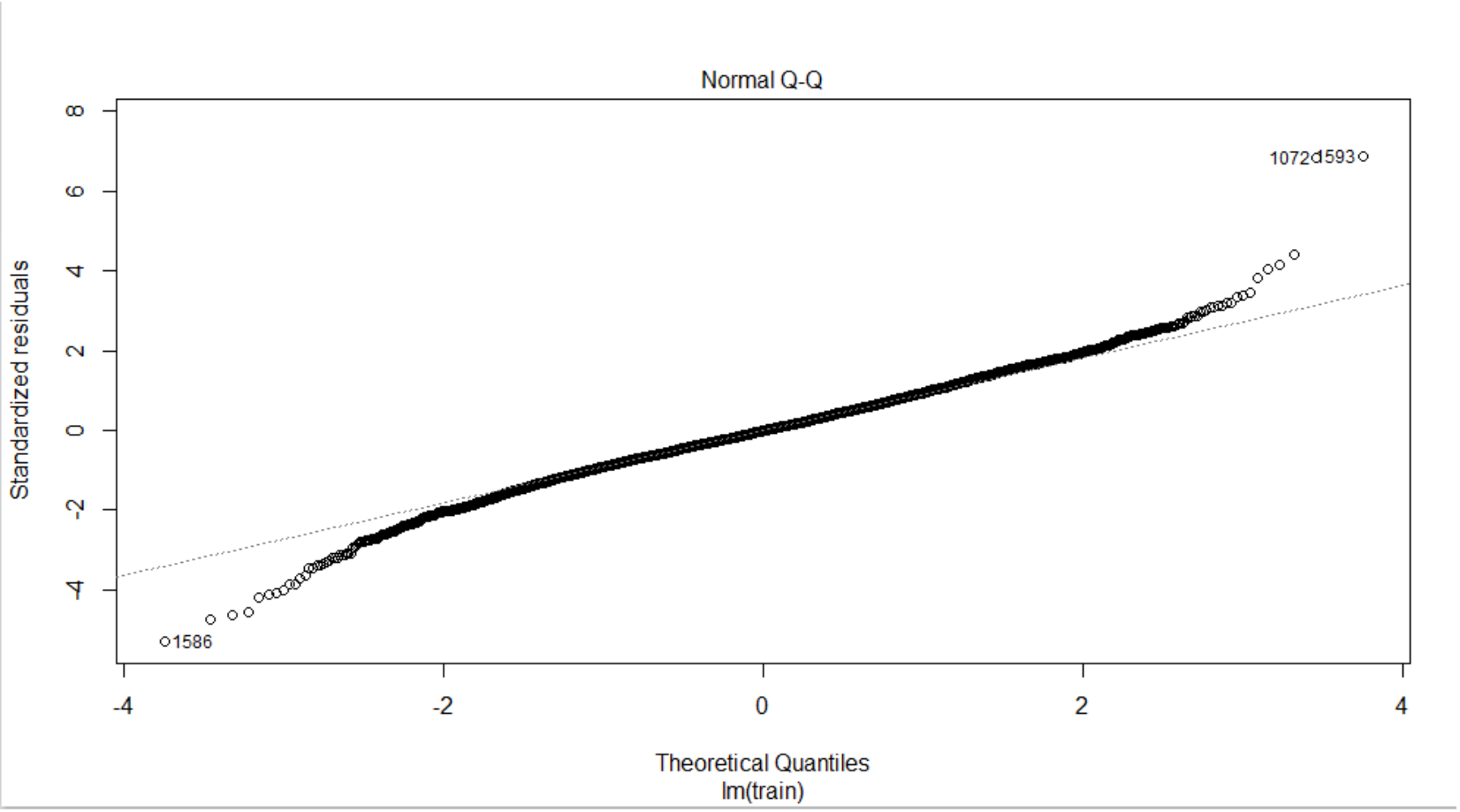
- Относится к машинному обучению и решается несколькими способами:
- Деревья решений
- Линейная регрессия
- Нейронные сети

Метрикой качества берем среднеквадратическое отклонение предсказаний на тестовом множестве.

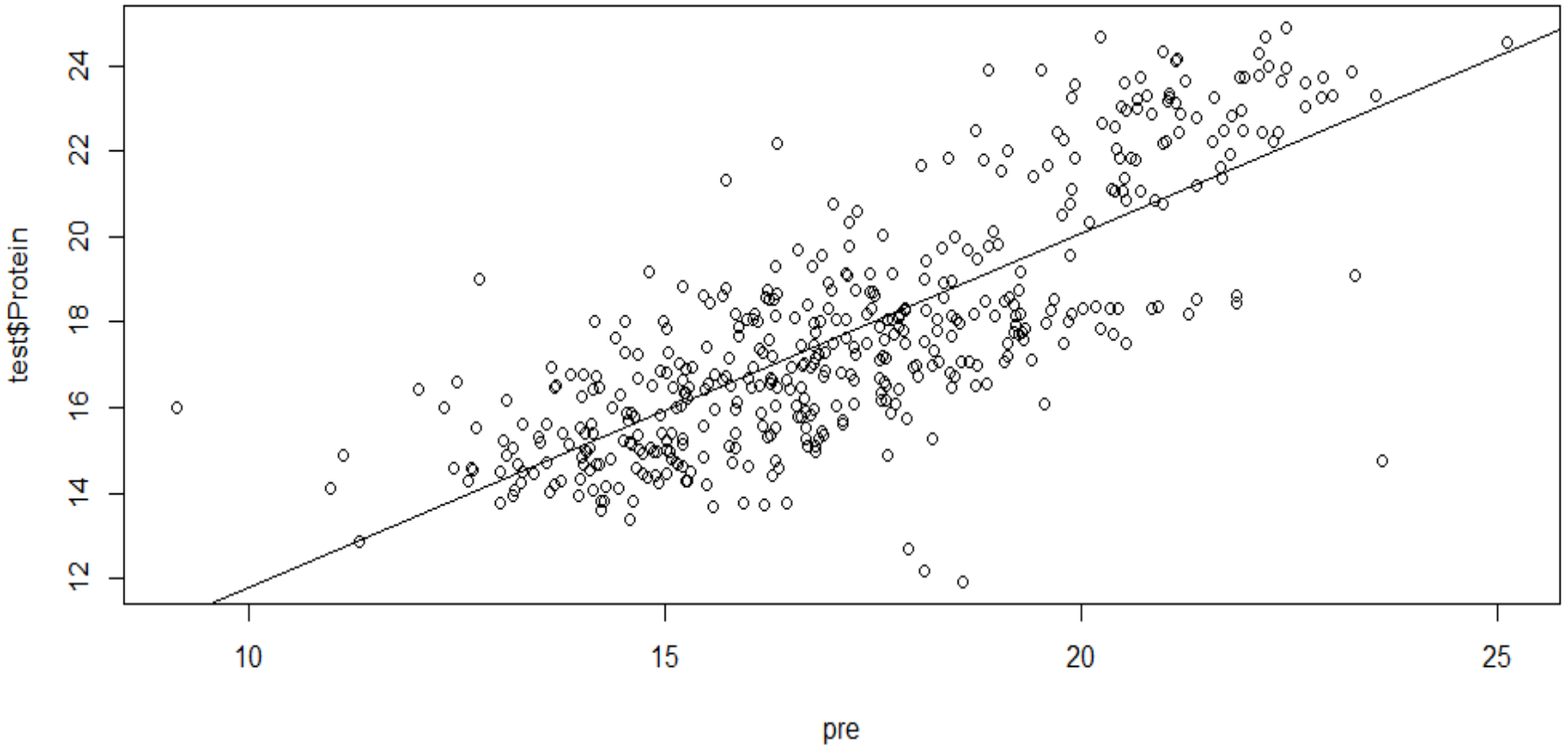
Методы

- Линейная регрессия после очистки данных дала результат 0.74 на тренировочных данных и 1.41 на тестовых.
- Деревья и ridge-регрессия не подошли
- Лучшим решением оказалась нейронная сеть,

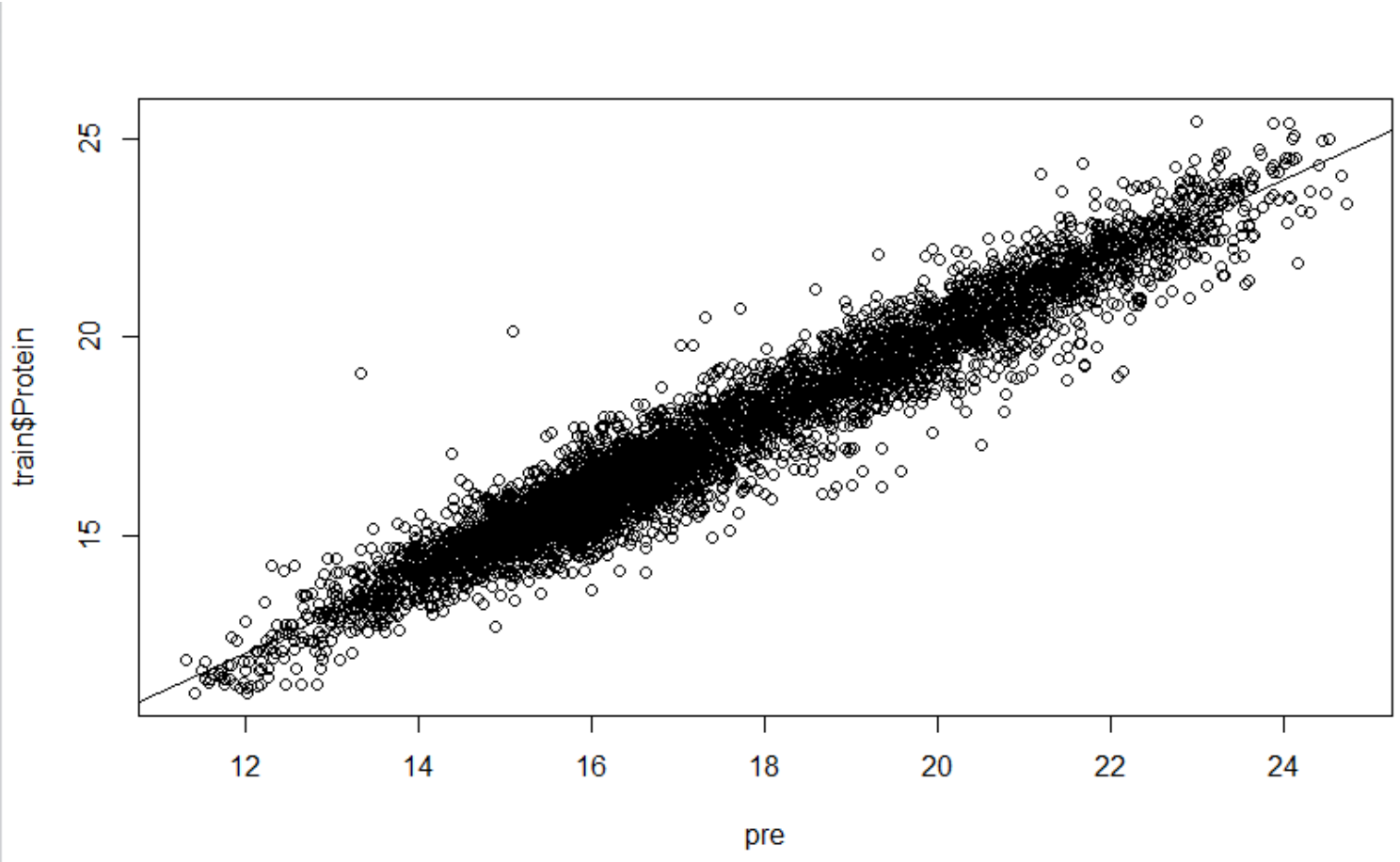

Residuals vs Fitted
1593
1072
1586
Residuals
6
4
2
0
-2
-4
10
15
20
25
Fitted values
lm(train)


Normal Q-Q
Standardized residuals
8
6
4
2
0
-2
-4
1072
593
1586
-4
-2
0
2
4
Theoretical Quantiles
lm(train)

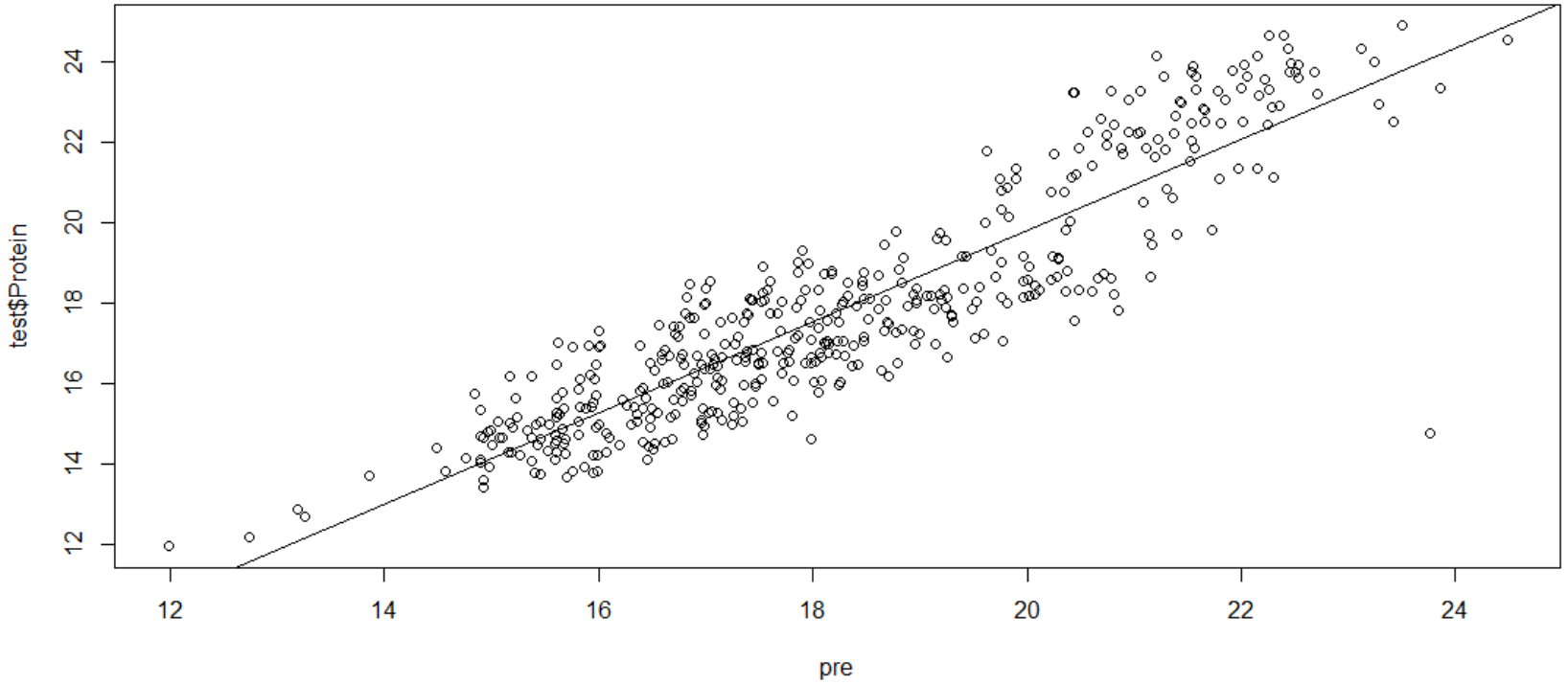
RMSEP 1.896502 – test(2.886988)
RMSEP 0.8085 – train(0.8241)



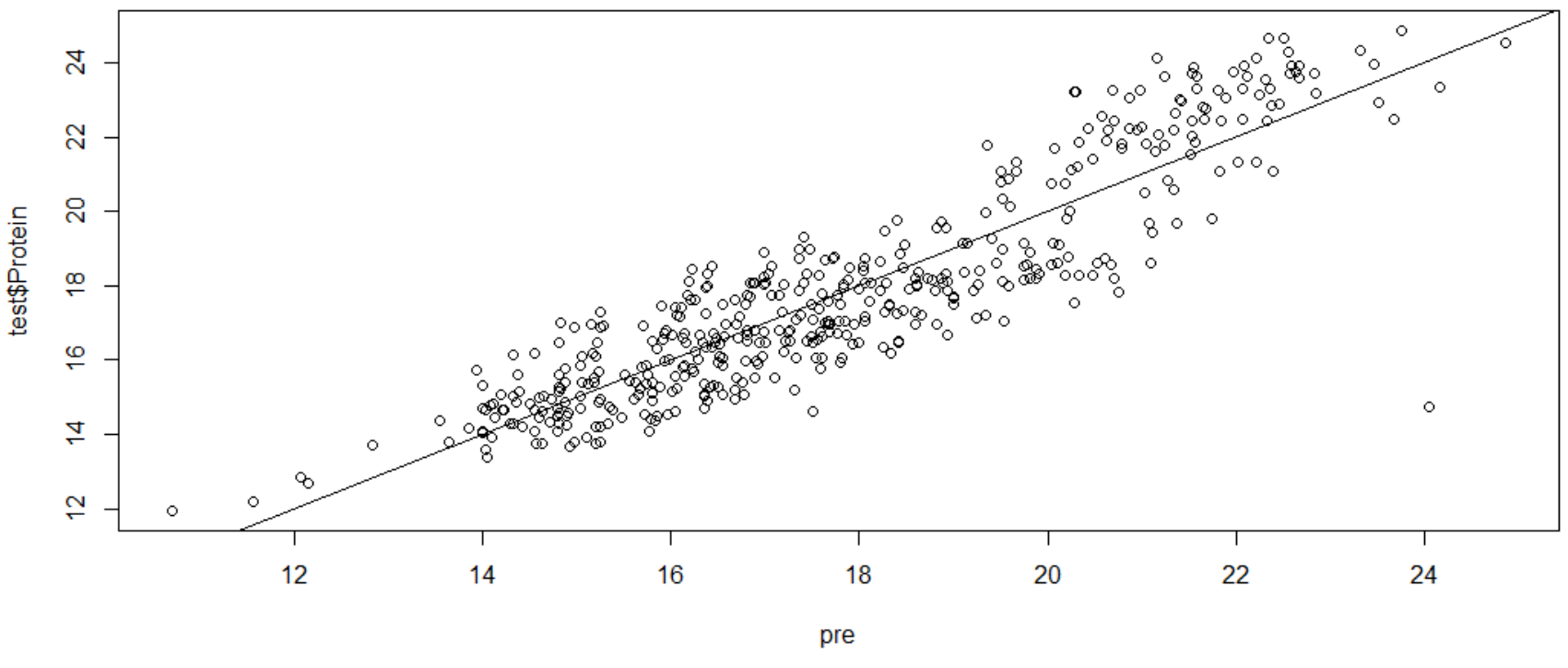
RMSEP 0.6179767 train (.8111362) NN



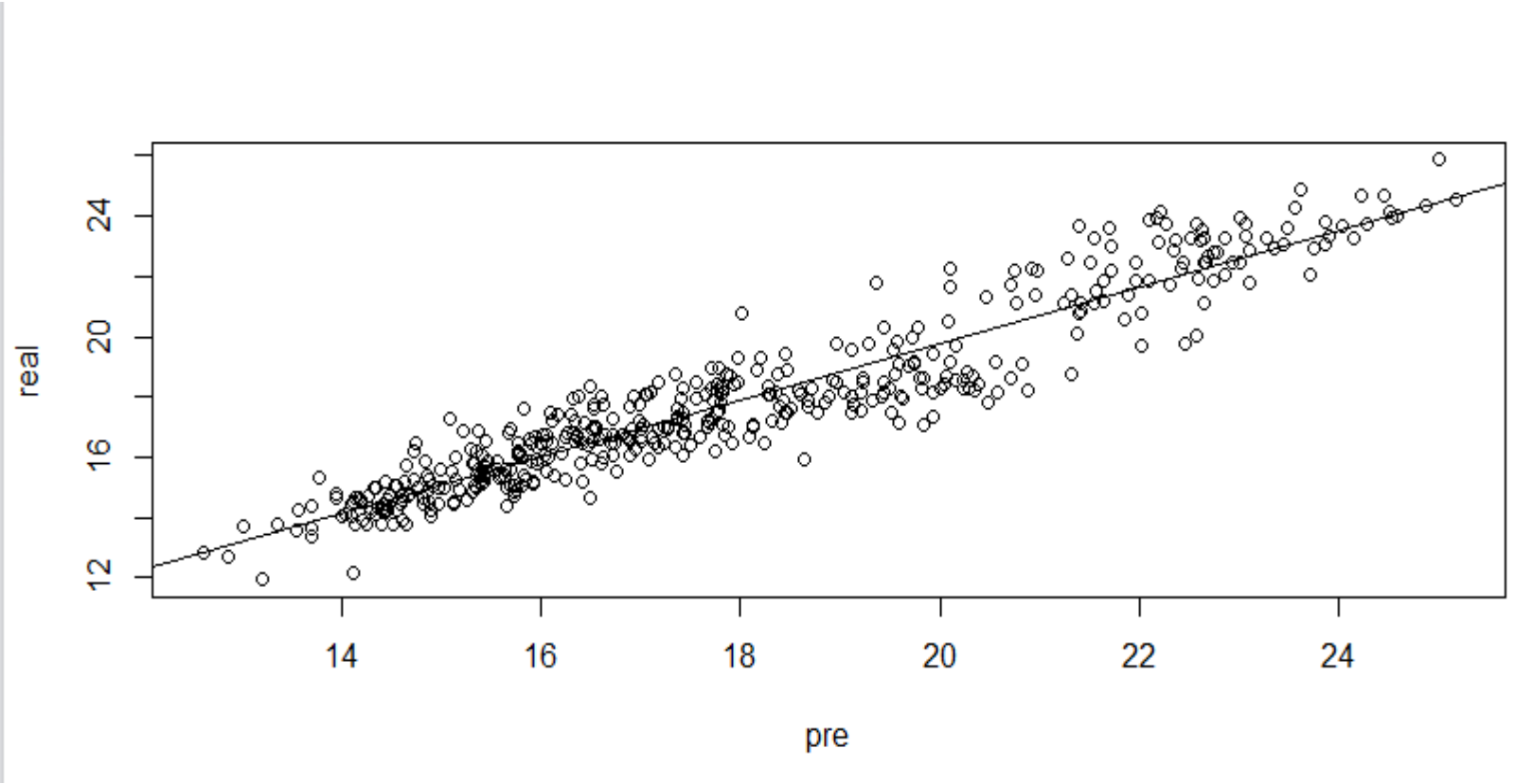
RMSEP 1.351587 (1.568266)

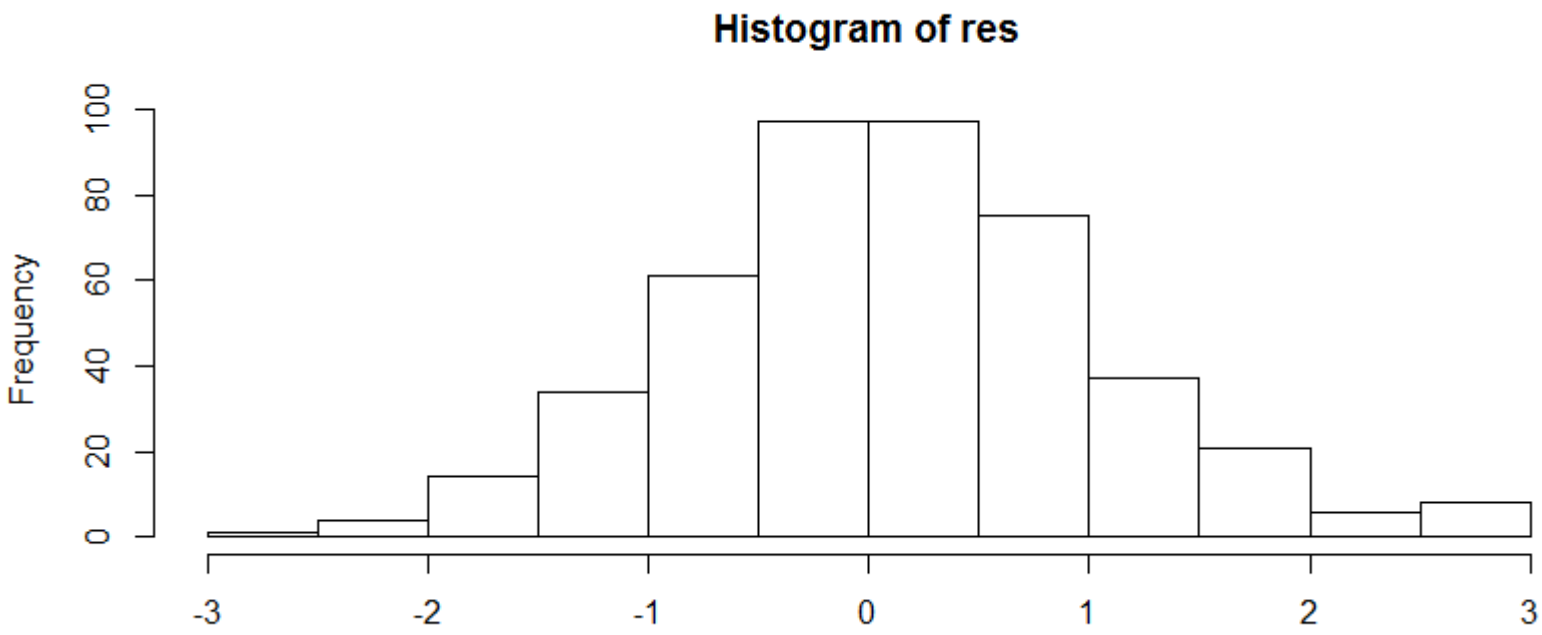


RMSEP 1.236491



RMSEP 0.95
4 нейрона, 198 итераций




Histogram of res
Frequency
0
20
40
60
80
100
-3
-2
-1
0
1
2
3