Прикладная статистика. Регрессионный анализ, пример решения задачи.

25 октября 2013 г.

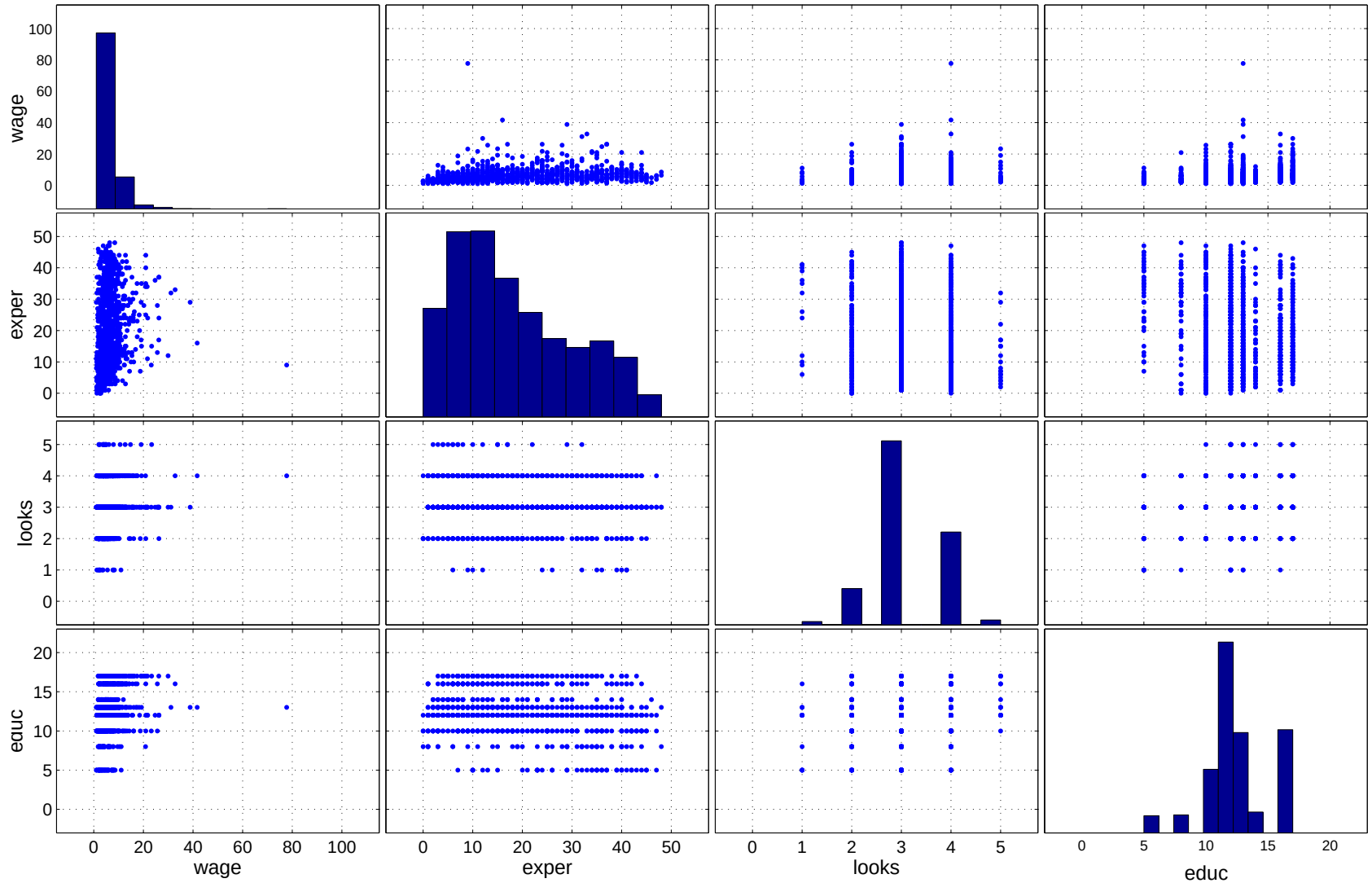
Влияние внешней привлекательности на уровень заработка

Hamermesh, D. S., and J. E. Biddle (1994), Beauty and the Labor Market, American Economic Review 84, 1174–1194: по 1260 опрошенным имеются следующие данные:

- заработная плата за час работы, \$,
- опыт работы, лет,
- образование, лет,
- внешняя привлекательность, в баллах от 1 до 5,
- бинарные признаки: пол, семейное положение, состояние здоровья (хорошее/плохое), членство в профсоюзе, цвет кожи (белый/чёрный), занятость в сере обслуживания (да/нет).

Оценить влияние внешней привлекательности на уровень заработка с учётом всех остальных факторов.

Данные


wage
exper
looks
educ
0 20 40 60 80 100
0 10 20 30 40 50
0 1 2 3 4 5
0 5 10 15 20

О необходимости визуализации данных

Пример:
http://en.wikipedia.org/wiki/Anscombe's_quartet

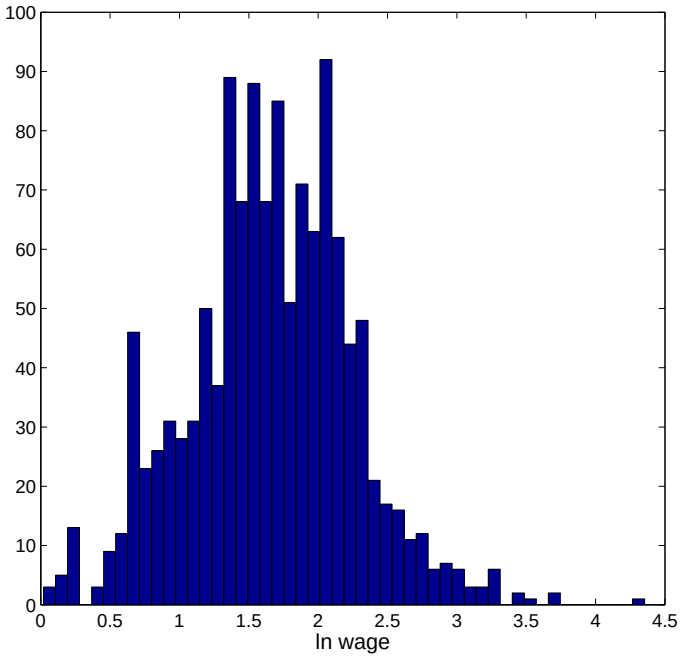
Данные

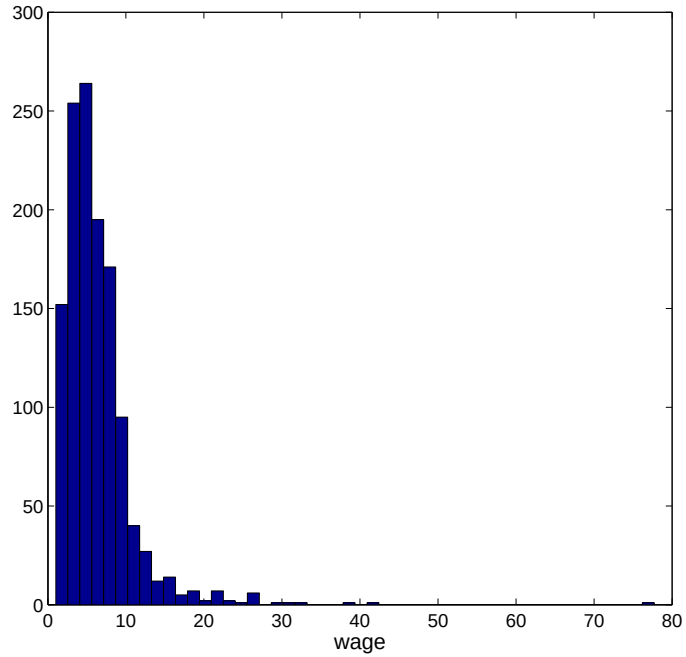
В группах $looks = 1$ и $looks = 5$ слишком мало наблюдений.

Превратим признак $looks$ в категориальный и закодируем при помощи фиктивных переменных:

$looks$	$aboveavg$	$belowavg$
< 3	1	0
3	0	0
> 3	0	1

Выбросы



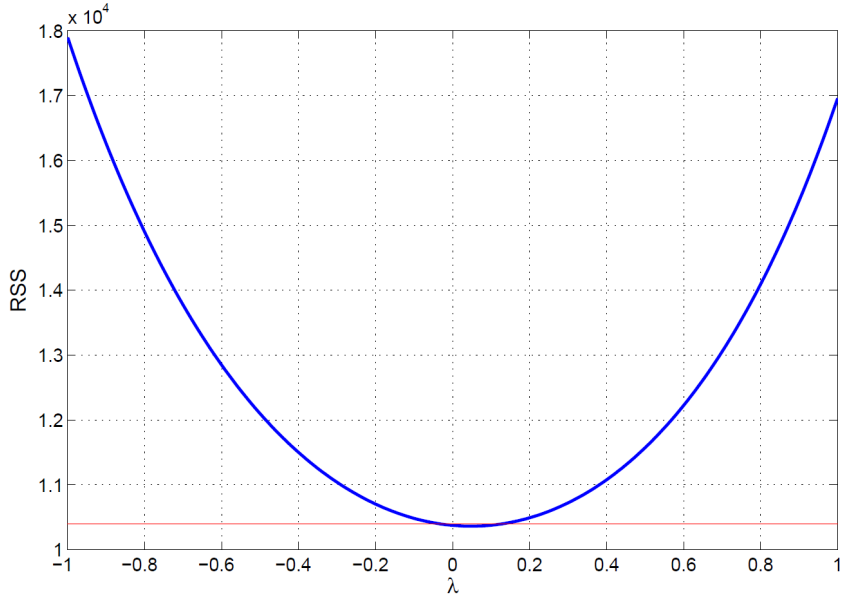


Больше 30 долларов в час в выборке получают только 5 человек. Исключим их.

Преобразование отклика

$$\frac{\max y}{\min y} = 29.4.$$

Найдём преобразование отклика при помощи метода Бокса-Кокса:



Доверительный интервал для λ определяется как пересечение кривой $RSS\left(\lambda\right)$ с линией уровня $\min\limits_{\lambda} RSS\left(\lambda\right) \cdot e^{\chi^2_{1,1-\alpha}/n}$.

95% доверительный интервал: $\left(-0.028, 0.124\right)$.

Возьмём $\lambda = 0$, т. е. будем делать регрессию логарифма отклика.

Модель 1

Построим линейную модель:

$$\begin{aligned}\ln wage = 0.43 + 0.01exper + 0.19union + 0.12goodhlth - 0.1black - \\ - 0.39female + 0.04married - 0.15service + 0.08educ - \\ - 0.004aboveavg - 0.13belogavg.\end{aligned}$$

$F = 78.63,\ p = 6 \times 10^{-125},\ R^2 = 0.387,\ R_a^2 = 0.382.$

Критерий	p-value
Шапиро-Уилка (нормальность)	1.0×10^{-4}
знаковых рангов (несмещённость)	0.8944
Бройша-Пагана (гомоскедастичность)	5.7×10^{-4}

Признаки, коэффициенты при которых значимо отличаются от нуля (множественная проверка): $exper$, $union$, $female$, $service$, $educ$, $belowavg$.

Модель 2

Редуцированная модель:

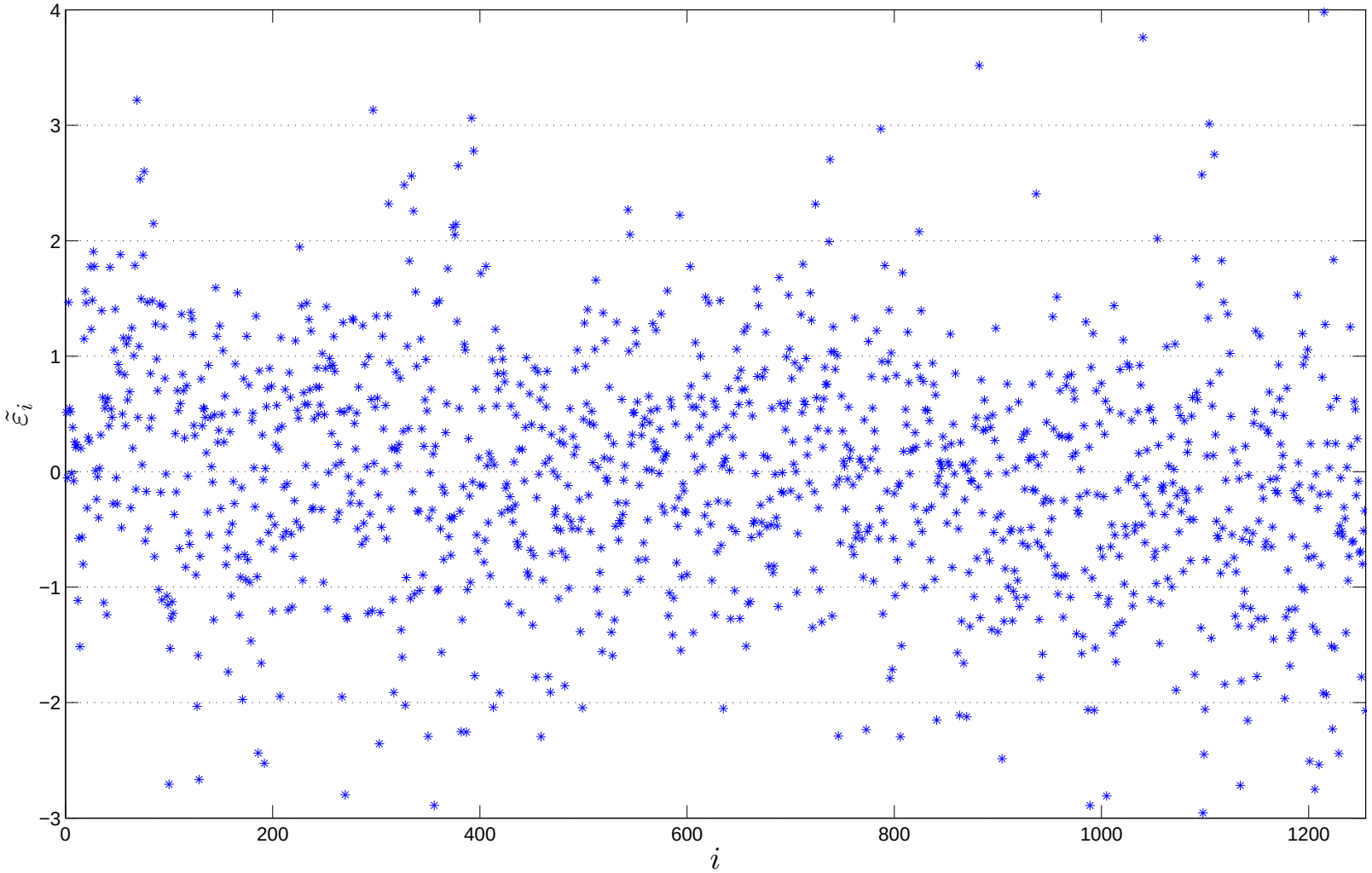
$$\ln wage = 0.54 + 0.01exper + 0.18union - 0.41female - 0.15service + \\ + 0.08educ - 0.008aboveavg - 0.12belogavg.$$

$F = 110.04,\ p = 1.5 \times 10^{-125},\ R^2 = 0.382,\ R_a^2 = 0.378.$

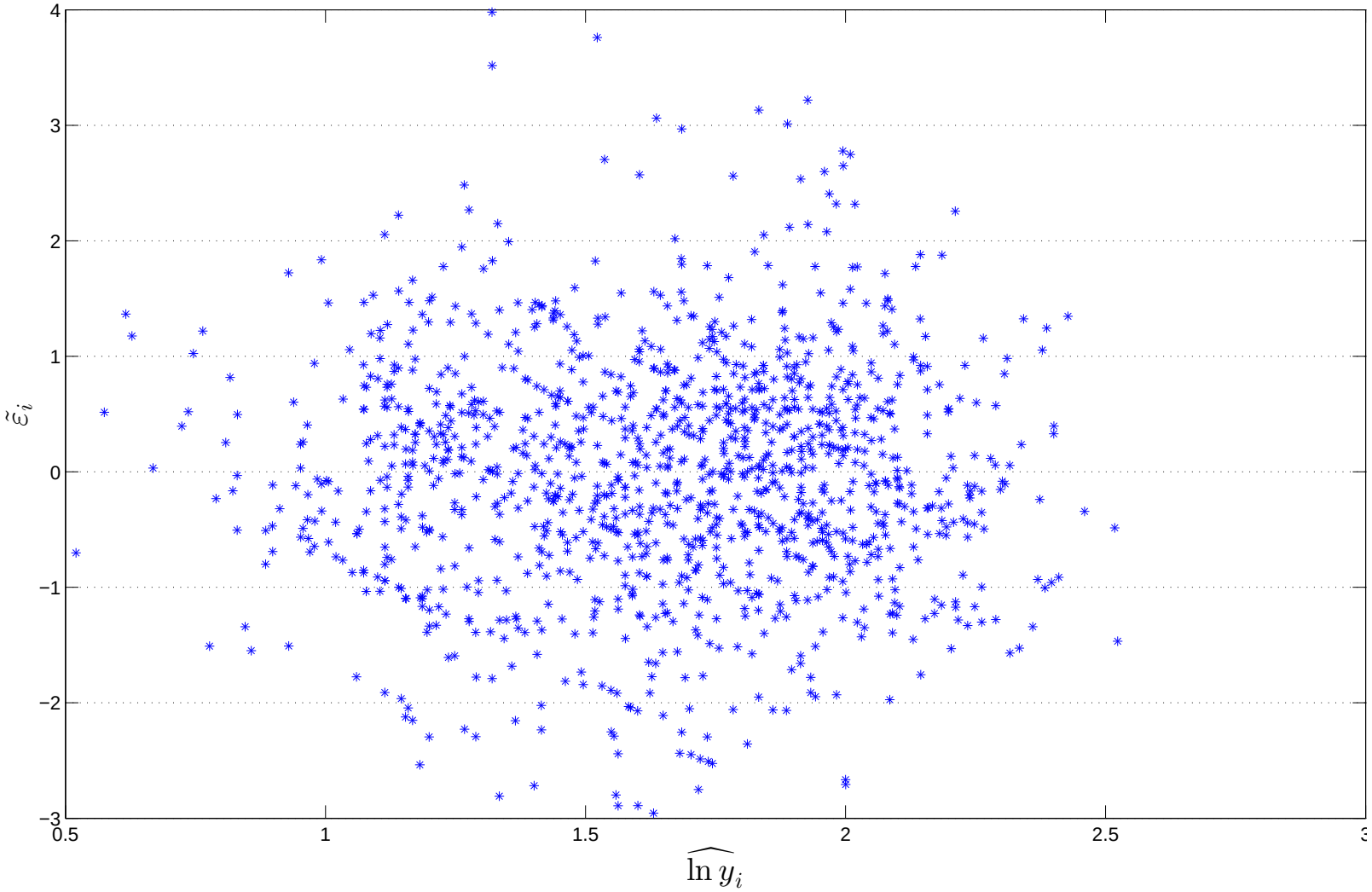
Критерий	p-value
Шапиро-Уилка (нормальность)	1.6×10^{-4}
знаковых рангов (несмещённость)	0.8480
Бройша-Пагана (гомоскедастичность)	4×10^{-5}

Значимы все признаки, кроме $aboveavg$.

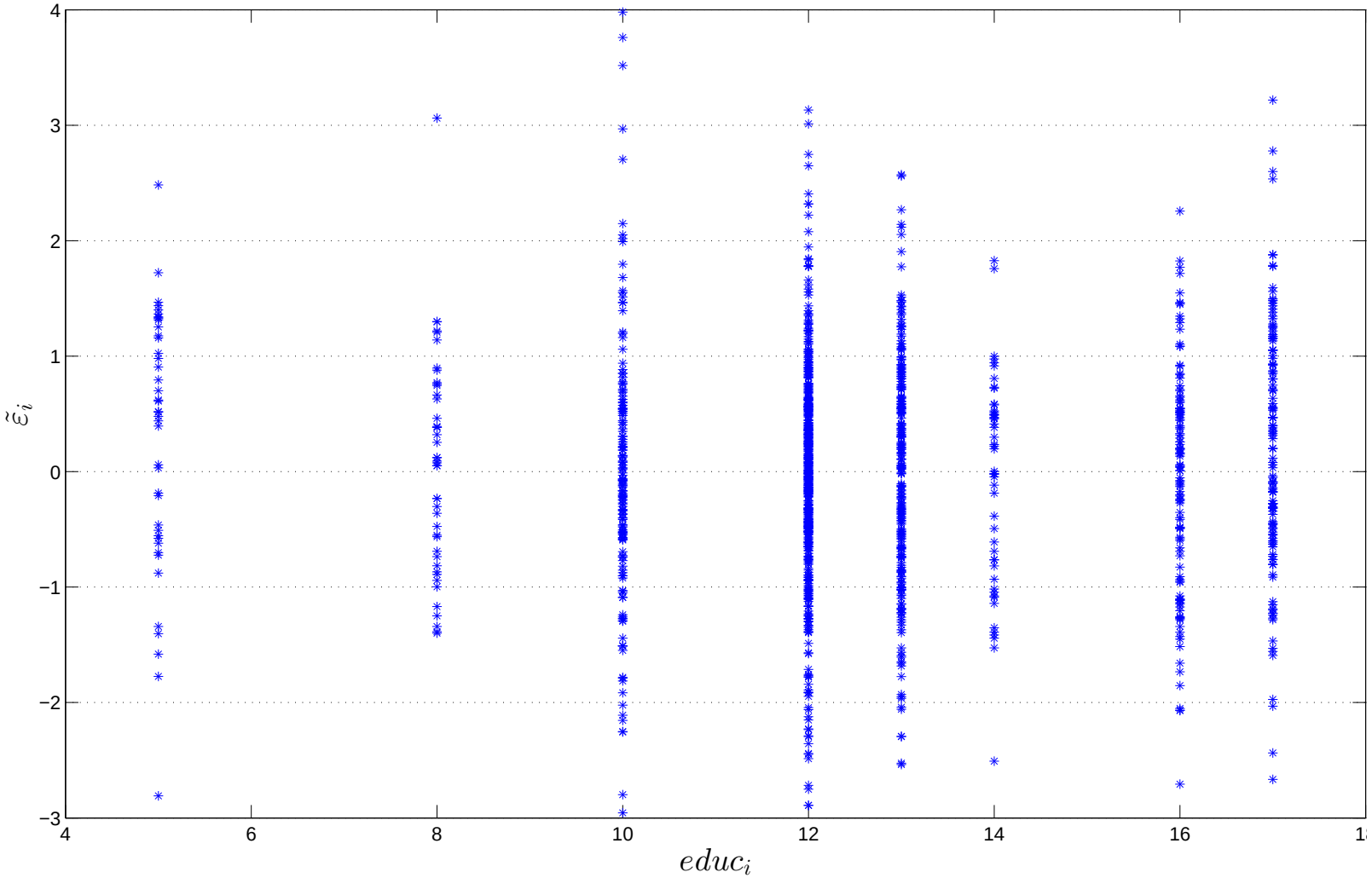
Остатки модели 2



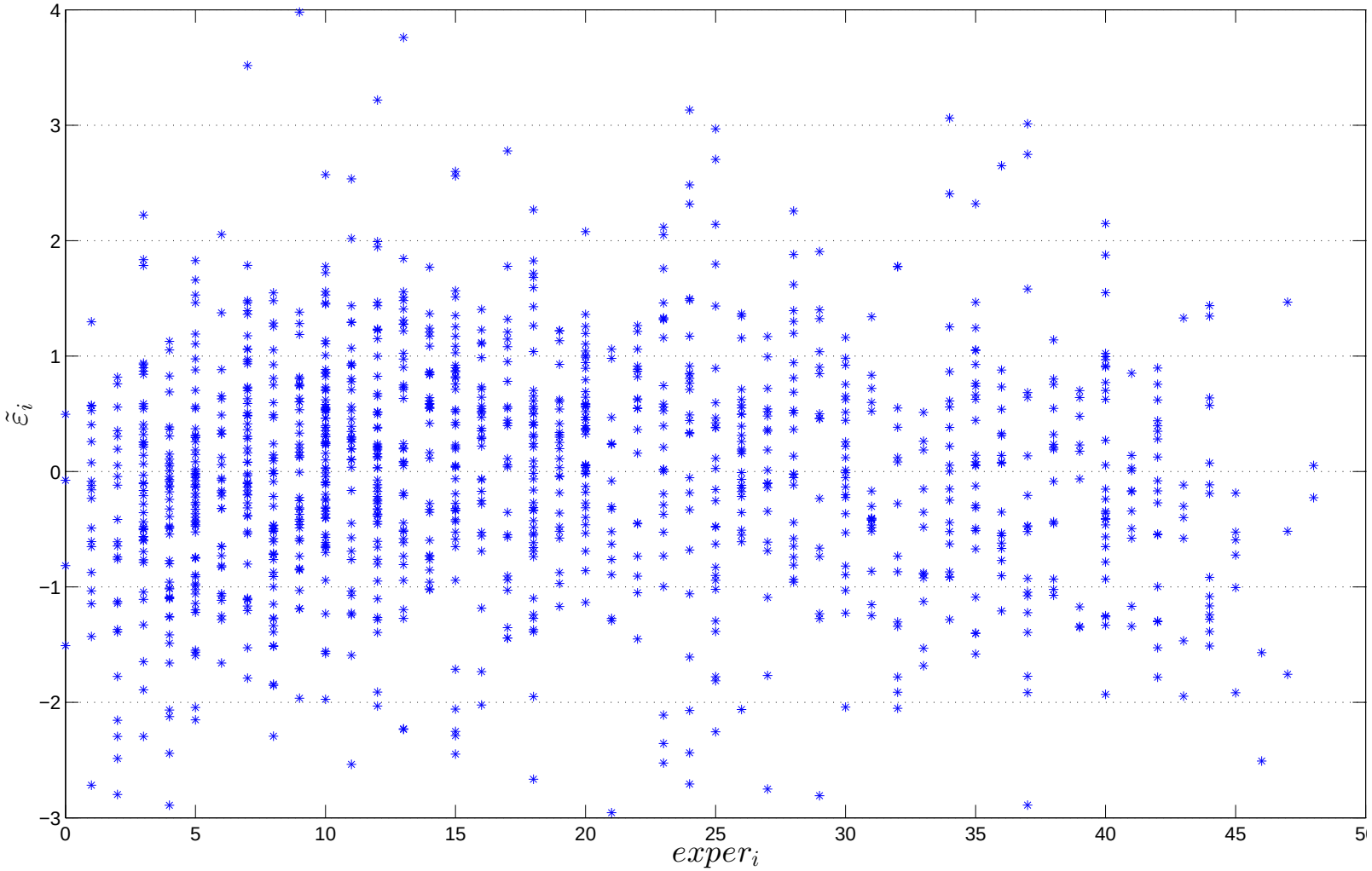
Остатки модели 2



Остатки модели 2



Остатки модели 2



Модель 3

Модель с квадратом признака $exper$:

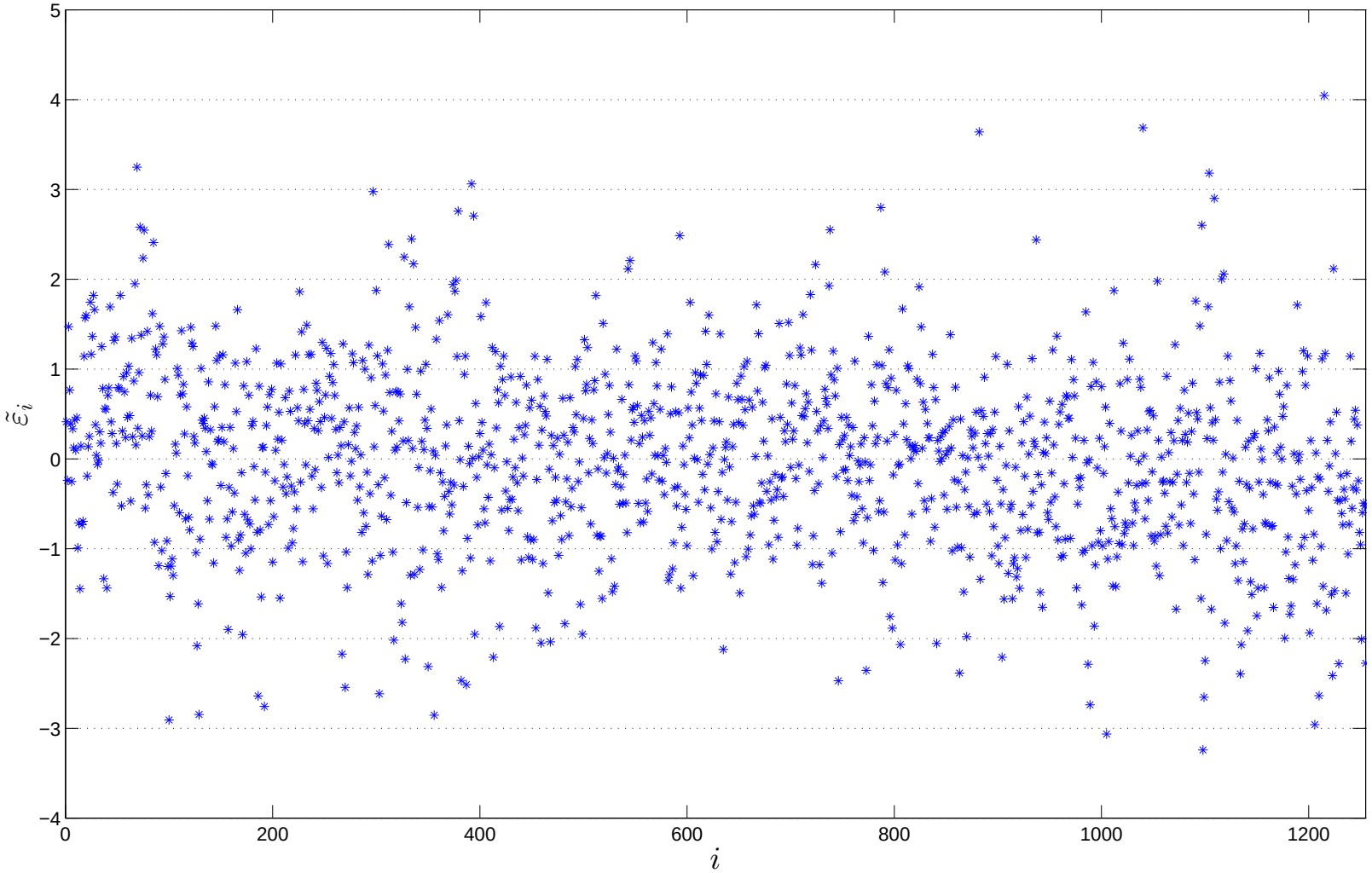
$$\ln wage = 0.4 + 0.04exper - 0.0006exper^2 + 0.18union - 0.4female - \\ - 0.16service + 0.08educ - 0.006aboveavg - 0.13belogavg.$$

$F = 104.92,\ p = 1.2 \times 10^{-133},\ R^2 = 0.402,\ R_a^2 = 0.399.$

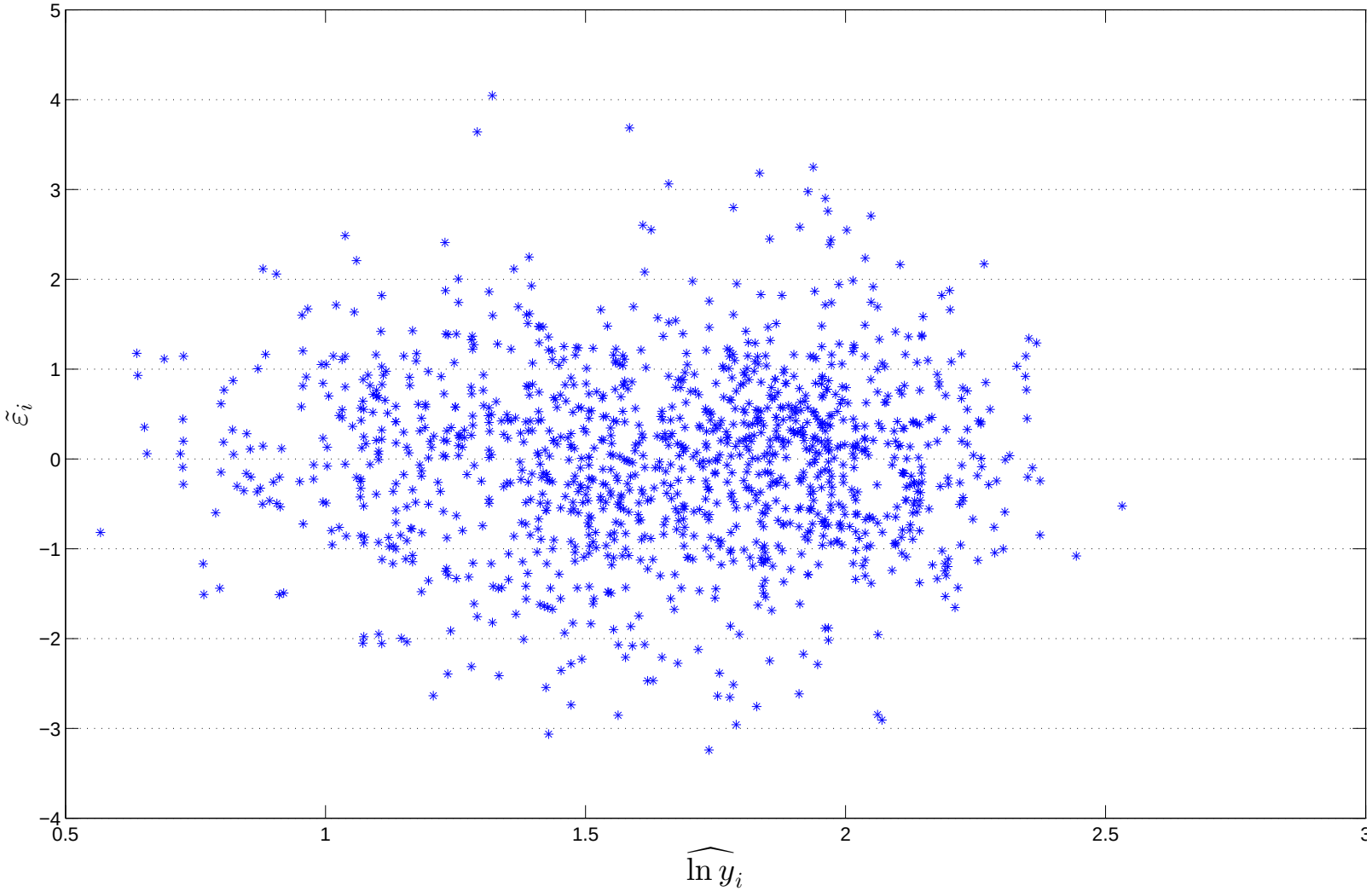
Критерий	p-value
Шапиро-Уилка (нормальность)	3.1×10^{-5}
знаковых рангов (несмещённость)	0.8571
Бройша-Пагана (гомоскедастичность)	5.5×10^{-6}

Значимы все признаки, кроме $aboveavg$.

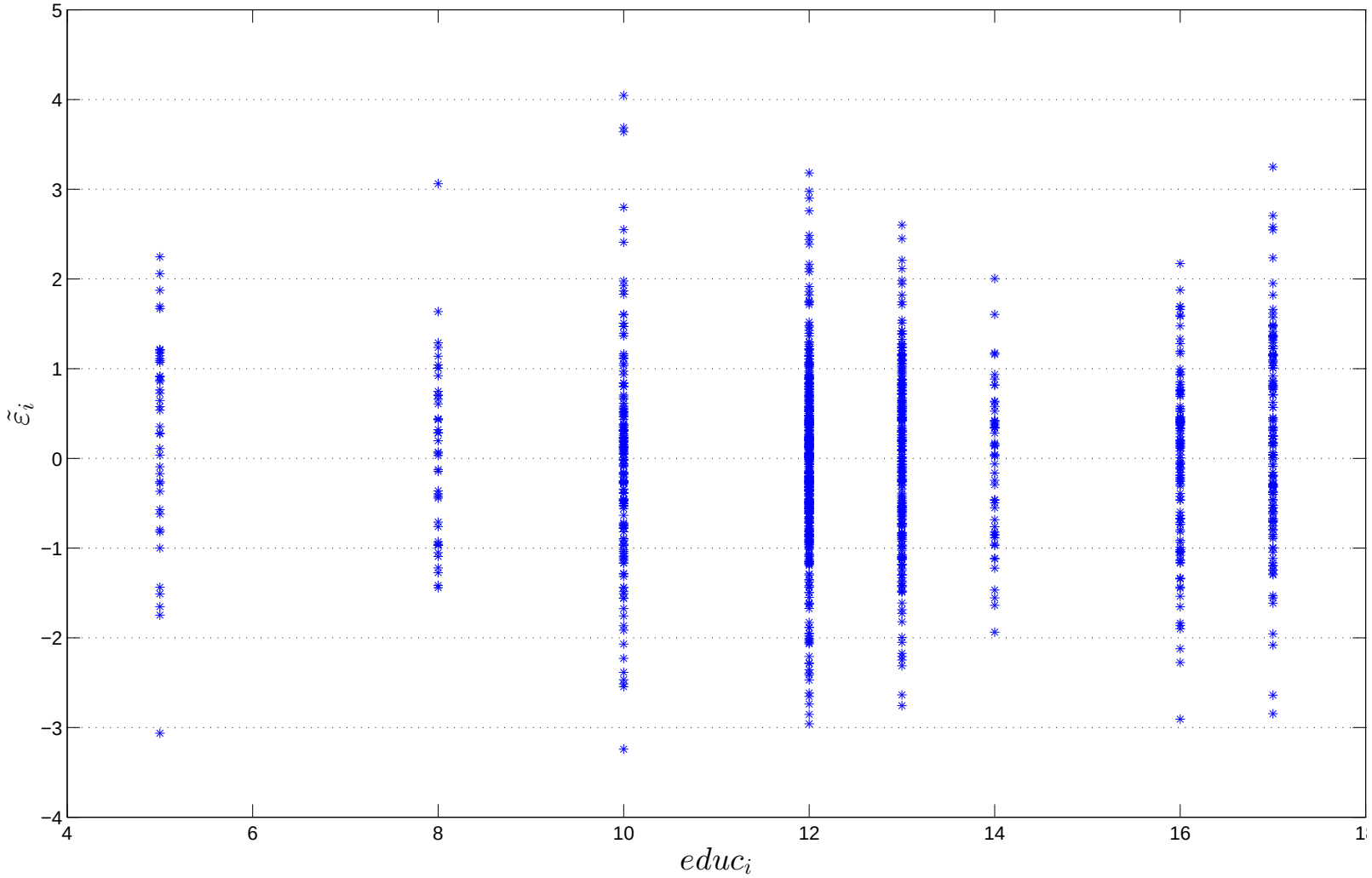
Остатки модели 3



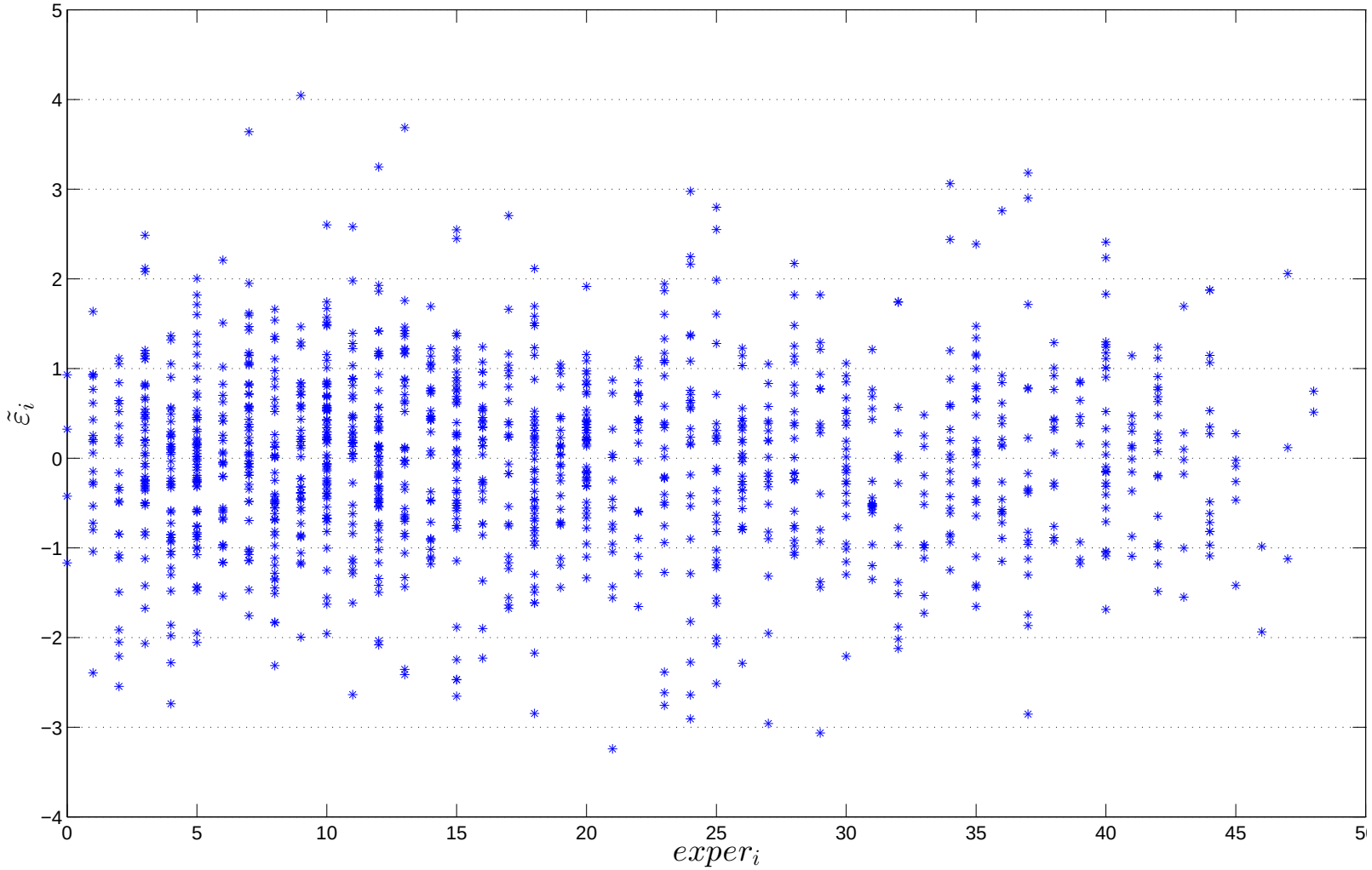
Остатки модели 3



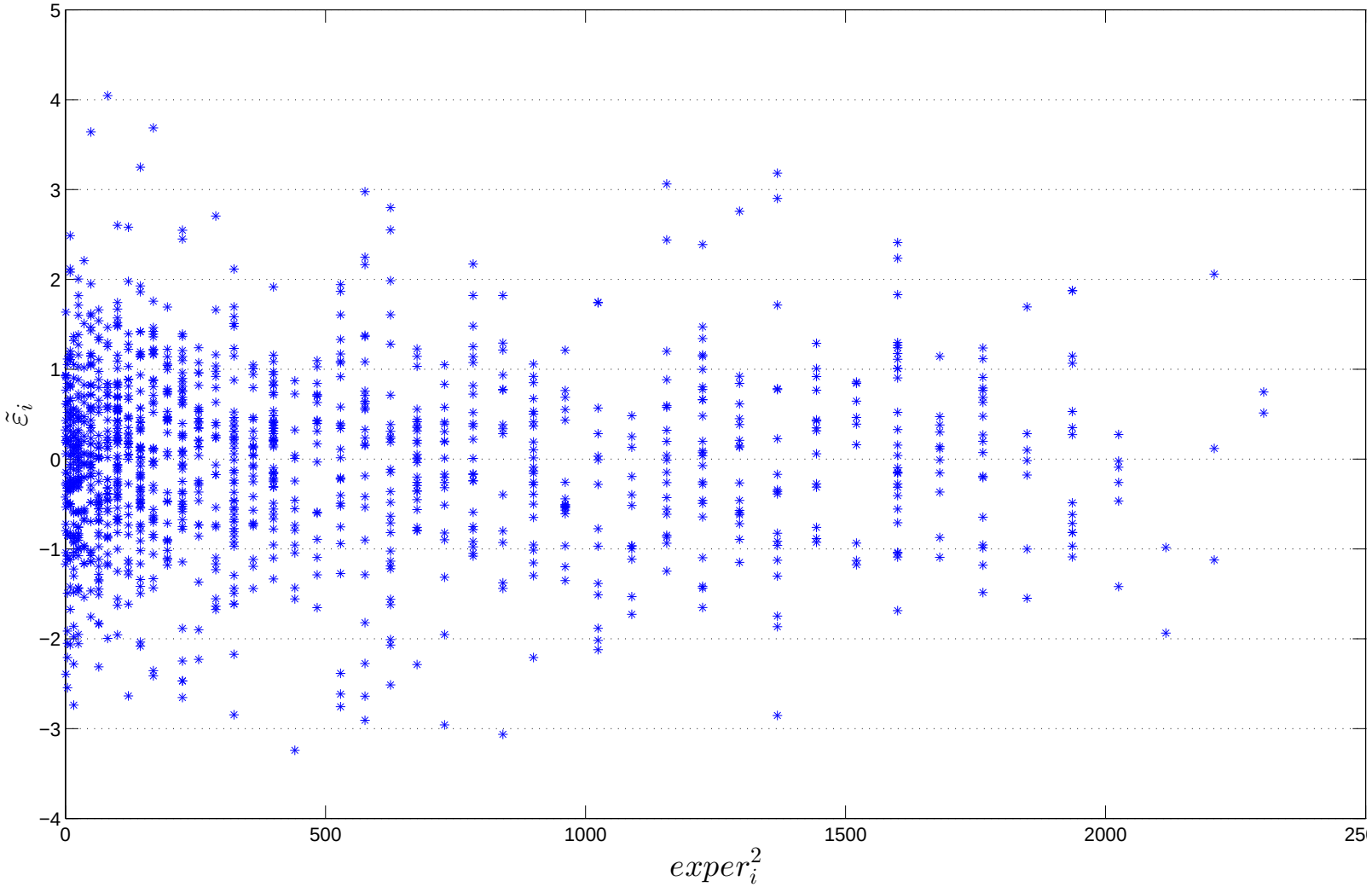
Остатки модели 3



Остатки модели 3



Остатки модели 3



Модель 4

Сделаем пошаговую регрессию со всеми попарными взаимодействиями и квадратами всех числовых признаков:

$$\begin{aligned}\ln wage = {} & 0.32 + 0.05exper + 0.27union - 0.25female + 0.06educ - \\ & - 0.005exper * union - 0.007exper * female - 0.008exper * service - \\ & - 0.31goodhlth * black + 0.01goodhlth * educ - 0.26goodhlth * belogavg + \\ & + 0.36black * female - 0.11female * married - 0.0007exper^2 - \\ & - 0.01aboveavg + 0.1belogavg.\end{aligned}$$

$F = 61.31,\ p = 1.4 \times 10^{-137},\ R^2 = 0.426,\ R_a^2 = 0.419.$

Критерий	p-value
Шапиро-Уилка (нормальность)	9.8×10^{-5}
знаковых рангов (несмещённость)	0.9204
Бройша-Пагана (гомоскедастичность)	5.1×10^{-5}

Модель 5

Чтобы упростить модель и повысить её интерпретируемость, исключим взаимодействия, для которых множественная проверка даёт достигаемый уровень значимости меньше 0.05:

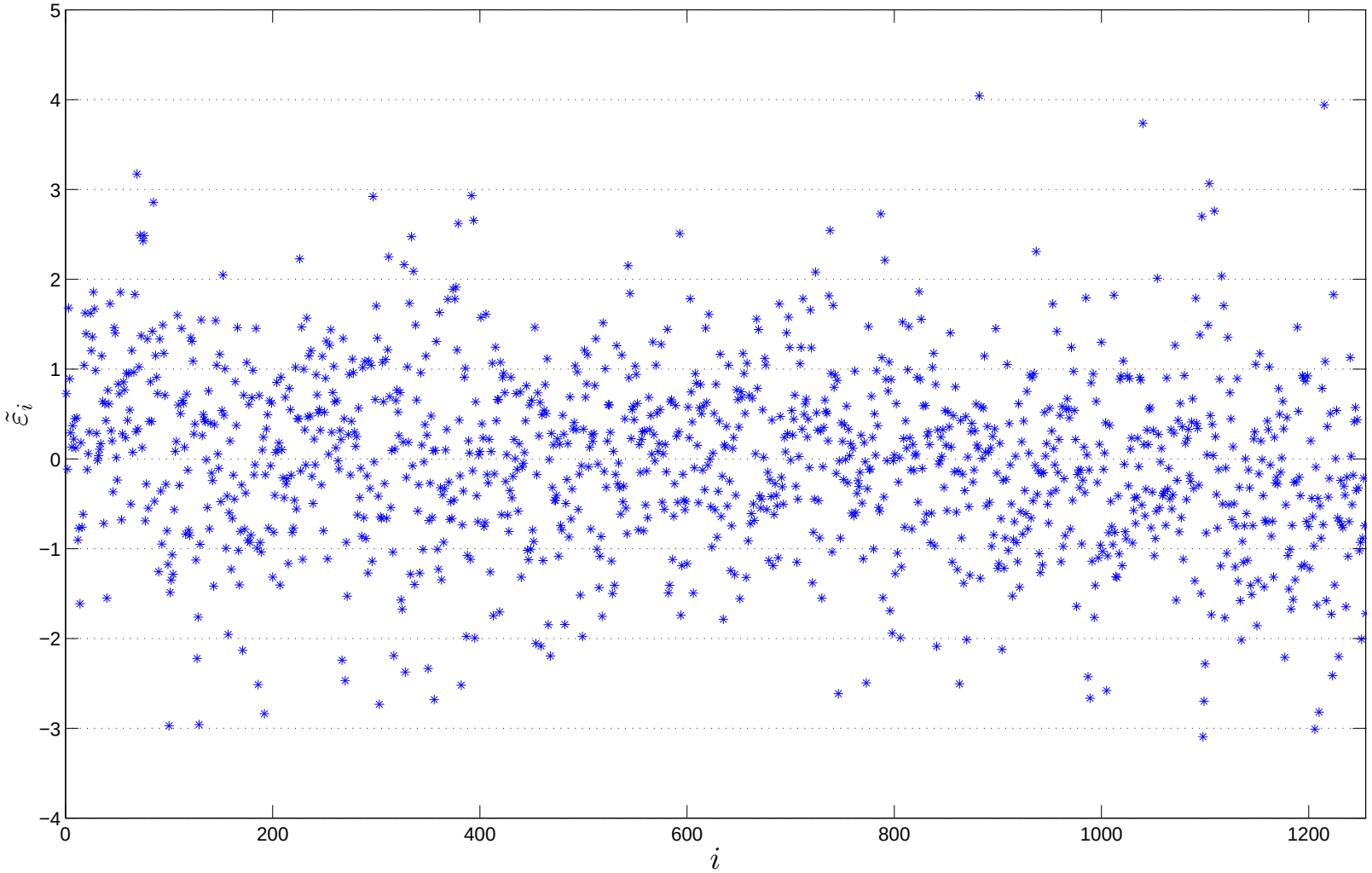
$$\begin{aligned}\ln wage = {} & 0.35 + 0.05exper + 0.19union - 0.32female + 0.06educ - \\ & - 0.007exper * female - 0.008exper * service - 0.3goodhlth * black + \\ & + 0.01goodhlth * educ + 0.36black * female - 0.0007exper^2 - \\ & - 0.008aboveavg - 0.14belogavg.\end{aligned}$$

$F = 74.90,\ p = 1.5 \times 10^{-137},\ R^2 = 0.420,\ R_a^2 = 0.414.$

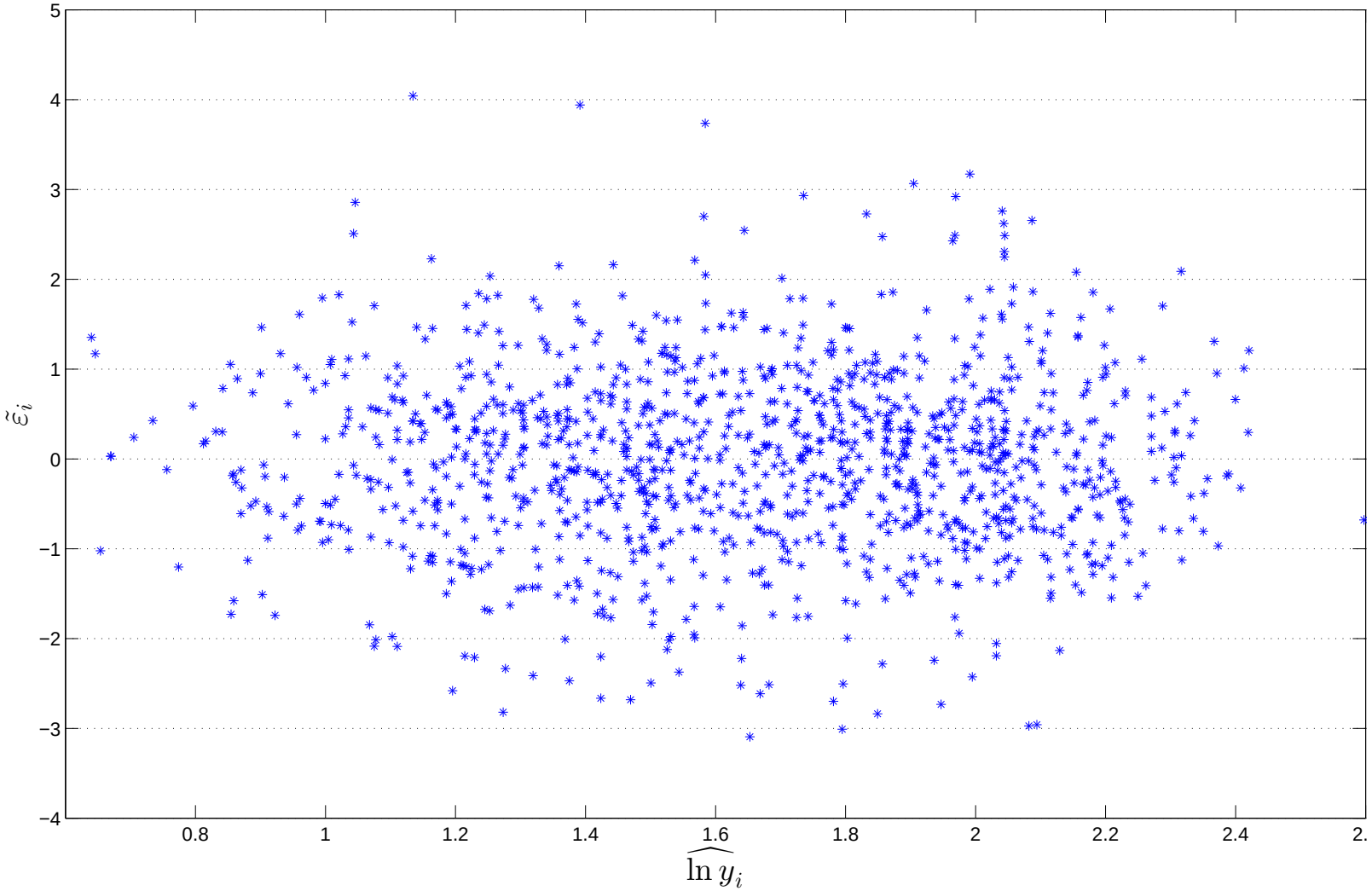
Критерий	p-value
Шапиро-Уилка (нормальность)	6.3×10^{-5}
знаковых рангов (несмещённость)	0.9360
Бройша-Пагана (гомоскедастичность)	2.1×10^{-4}

Критерий Давидсона-МакКинли показывает превосходство модели 5 над моделью 3 ($p_1 = 7.1 \times 10^{-10},\ p_2 = 0.0851$).

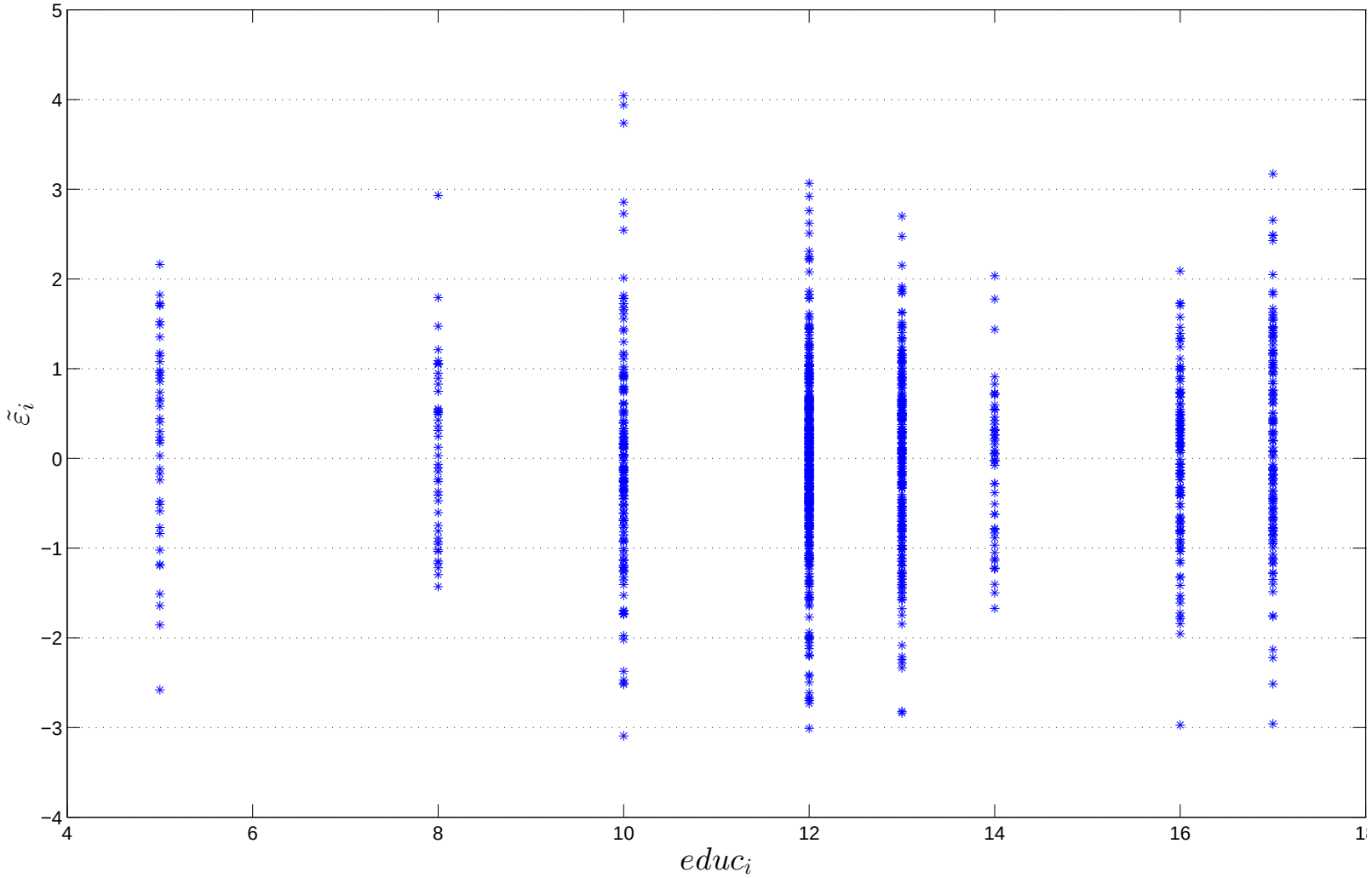
Остатки модели 5



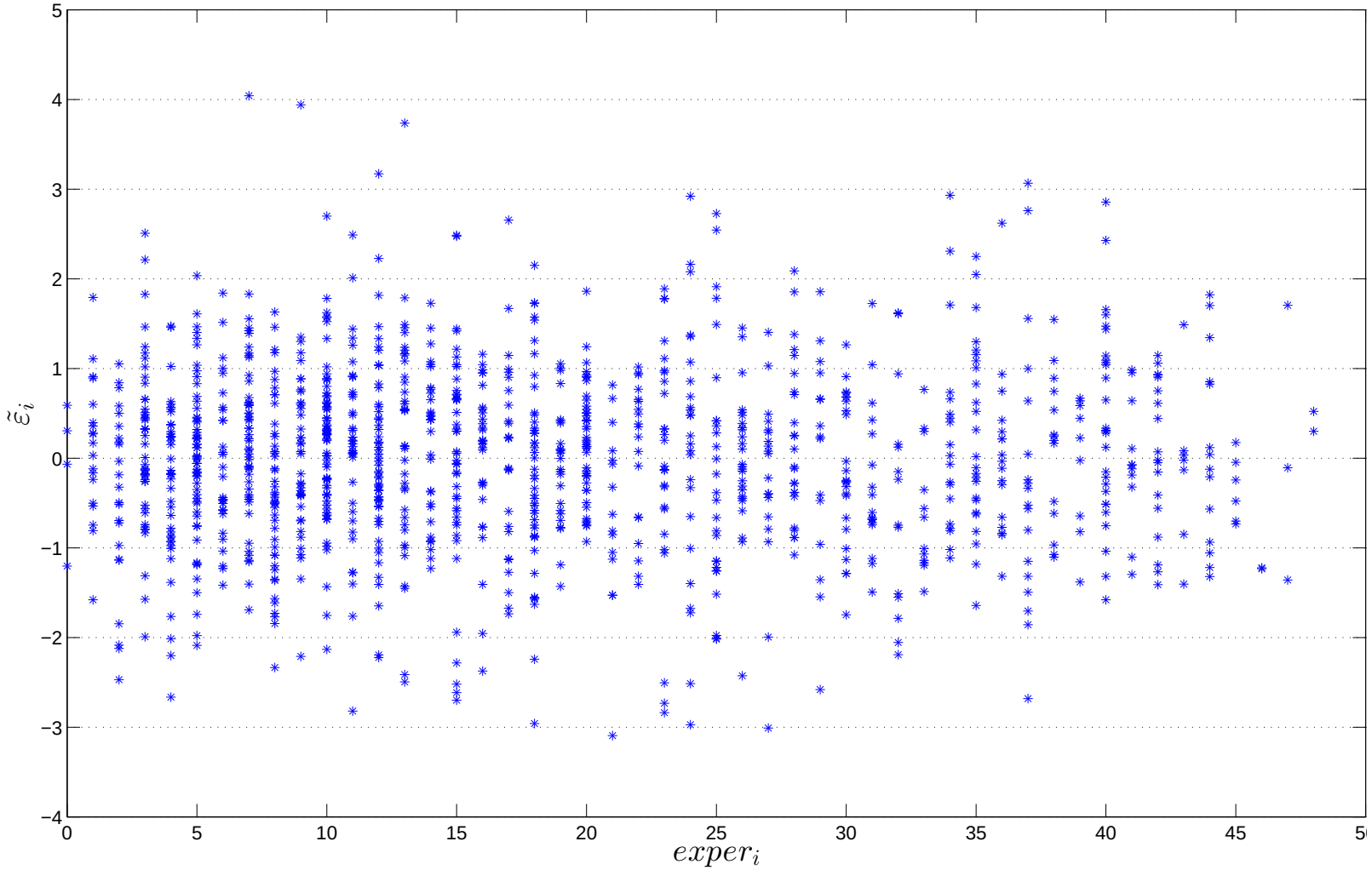
Остатки модели 5



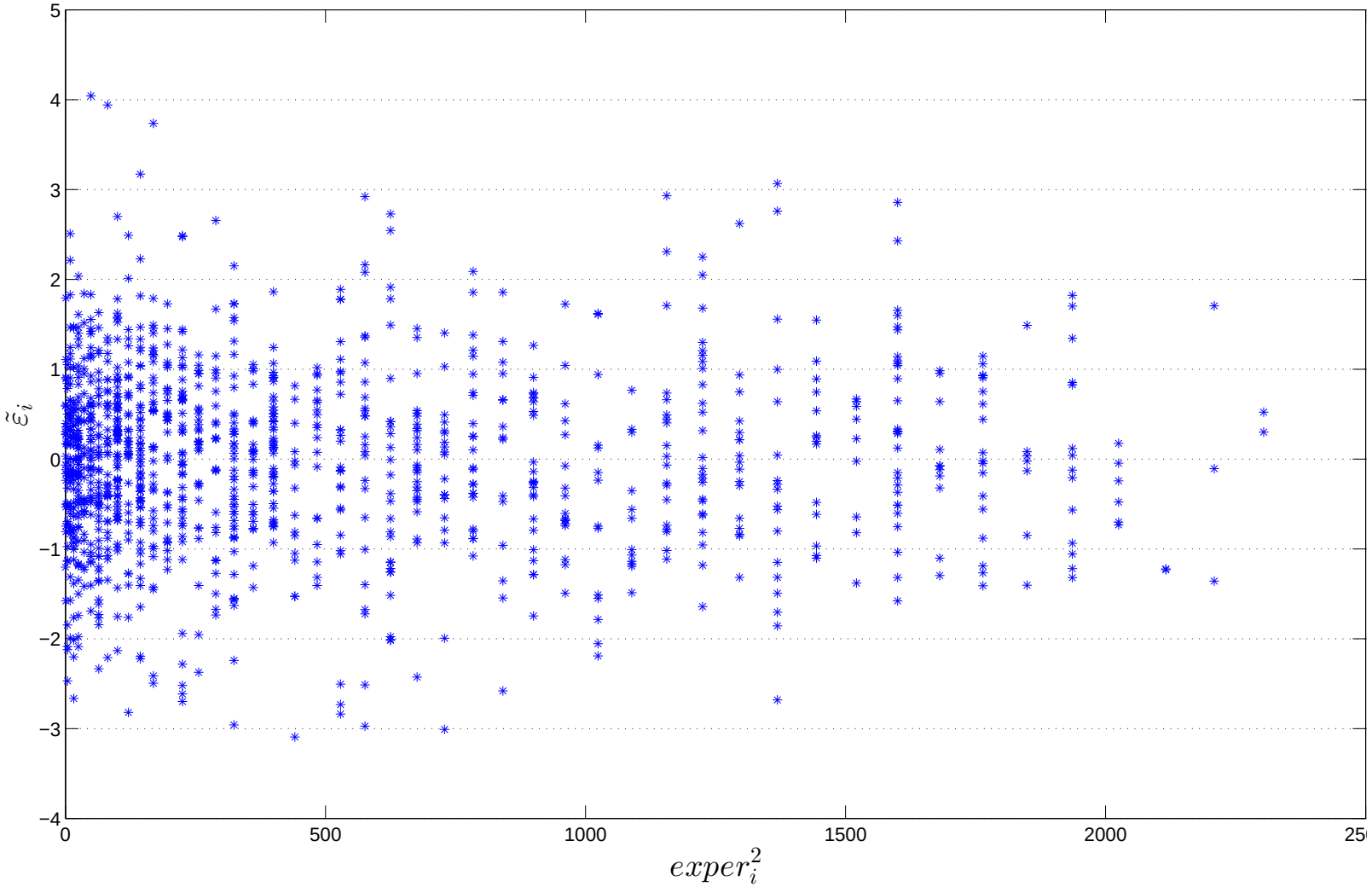
Остатки модели 5



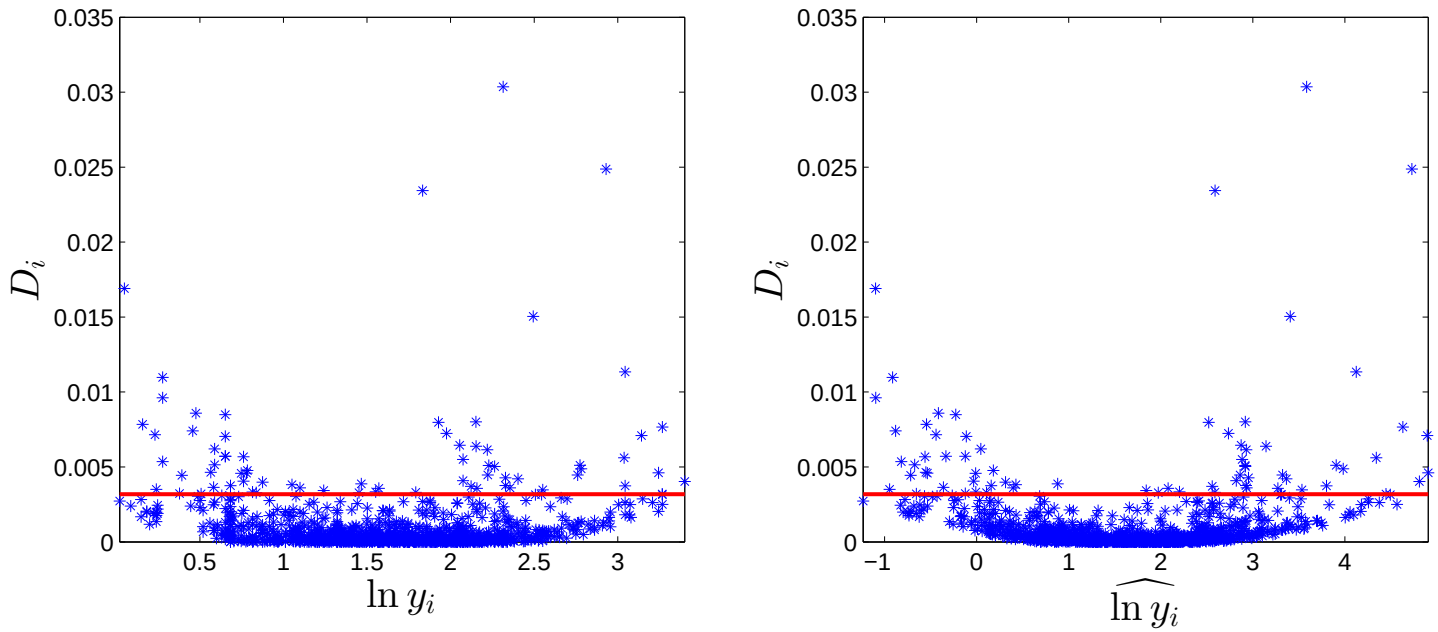
Остатки модели 5



Остатки модели 5



Расстояние Кука для модели 5



Порог $4/n$ расстояние Кука превышает для 70 наблюдений (из них 14 $aboveavg$ и 12 $belowavg$). Исключим их.

Модель 6

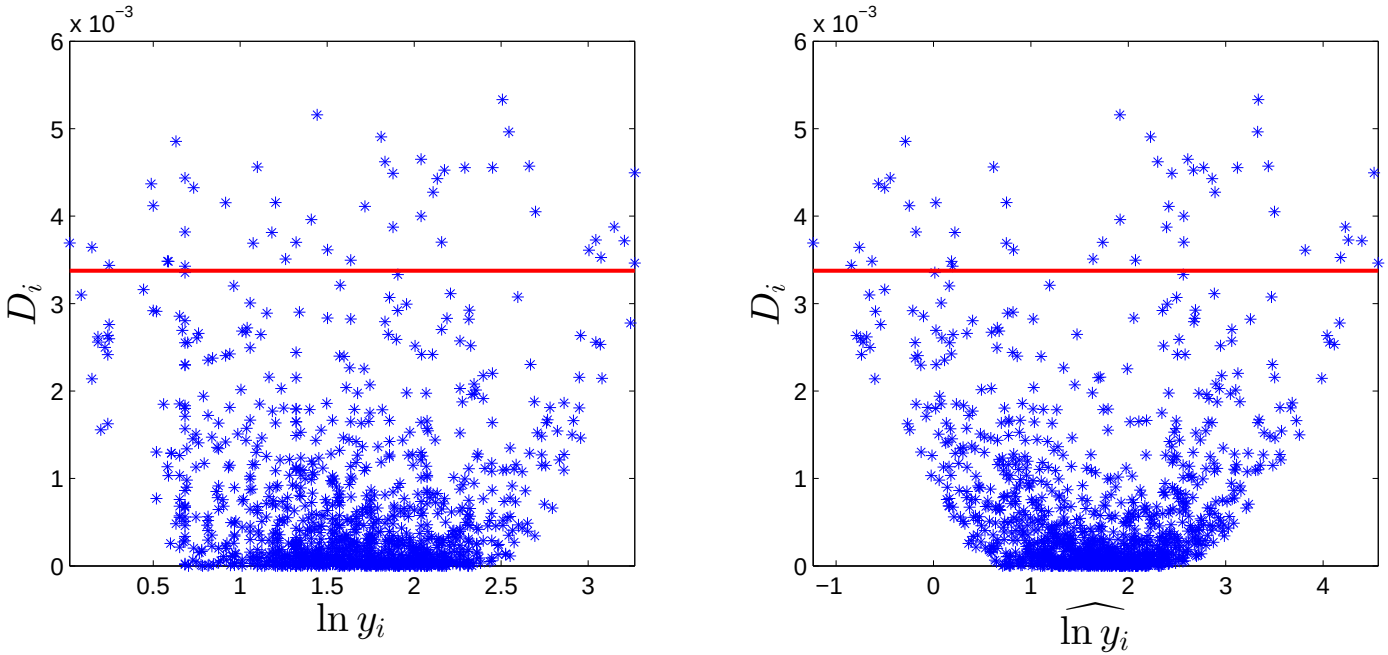
Модель, настроенная на оставшихся наблюдениях:

$$\begin{aligned}\ln wage = {} & 0.28 + 0.05exper + 0.20union - 0.31female + 0.06educ - \\ & - 0.007exper * female - 0.007exper * service - 0.26goodhlth * black + \\ & + 0.01goodhlth * educ + 0.30black * female - 0.0008exper^2 + \\ & + 0.009aboveavg - 0.15belogavg.\end{aligned}$$

$F = 94.52,\ p = 9.3 \times 10^{-163},\ R^2 = 0.492,\ R_a^2 = 0.487.$

Критерий	p-value
Шапиро-Уилка (нормальность)	0.1174
знаковых рангов (несмещённость)	≈ 1
Бройша-Пагана (гомоскедастичность)	8.0×10^{-5}

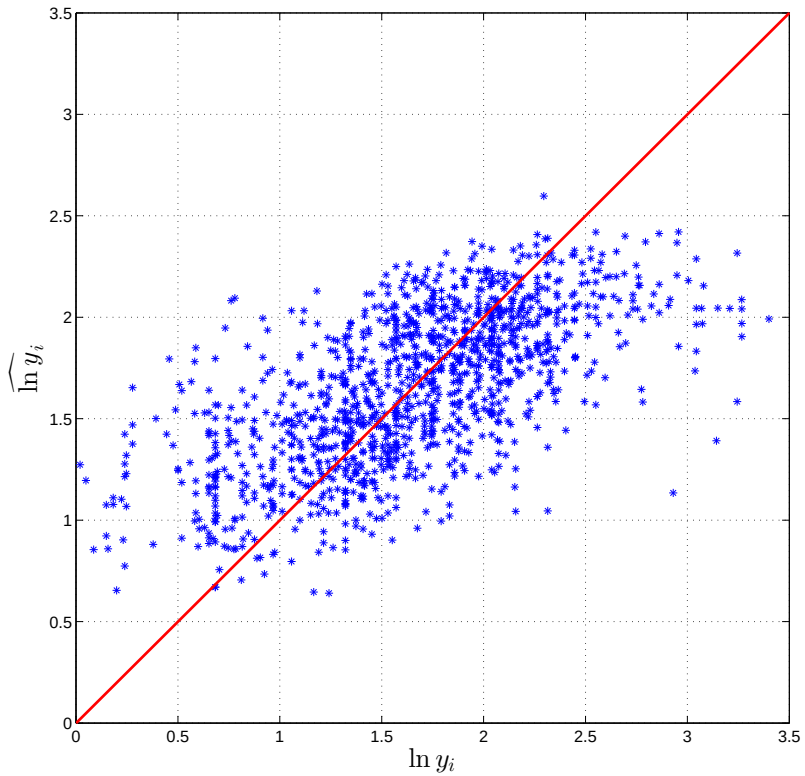
Расстояние Кука для модели 6



Порог $4/n$ расстояние Кука превышает для 47 наблюдений (из них 13 $aboveavg$ и 10 $belowavg$).

Результат

Итоговая модель (№5) объясняет 42% вариации логарифма отклика:



С учётом дополнительных факторов, участники опроса с привлекательностью ниже среднего получают на 13% меньше (95% доверительный интервал $(5.7\%, 19.2\%)$), а с привлекательностью выше среднего — на 0.8% меньше (95% доверительный интервал $(-5.0\%, 6.2\%)$).

Требования к решению задачи методом линейной регрессии

- визуализация данных, анализ распределения признаков (оценка необходимости трансформации), оценка наличия выбросов;
- оценка необходимости преобразования отклика и его поиск методом Бокса-Кокса;
- отбор признаков;
- визуальный анализ остатков;
- проверка гипотез об остатках: нормальность, несмещённость, гомоскедастичность;
- анализ необходимости добавления взаимодействий и квадратов признаков;
- расчёт расстояний Кука, возможное удаление выбросов, обновление модели;
- выводы.

Прикладная статистика
Регрессионный анализ, пример решения задачи.

Рябенко Евгений
riabenko.e@gmail.com