Прикладная статистика. Регрессионный анализ, пример решения задачи.

15 апреля 2013 г.

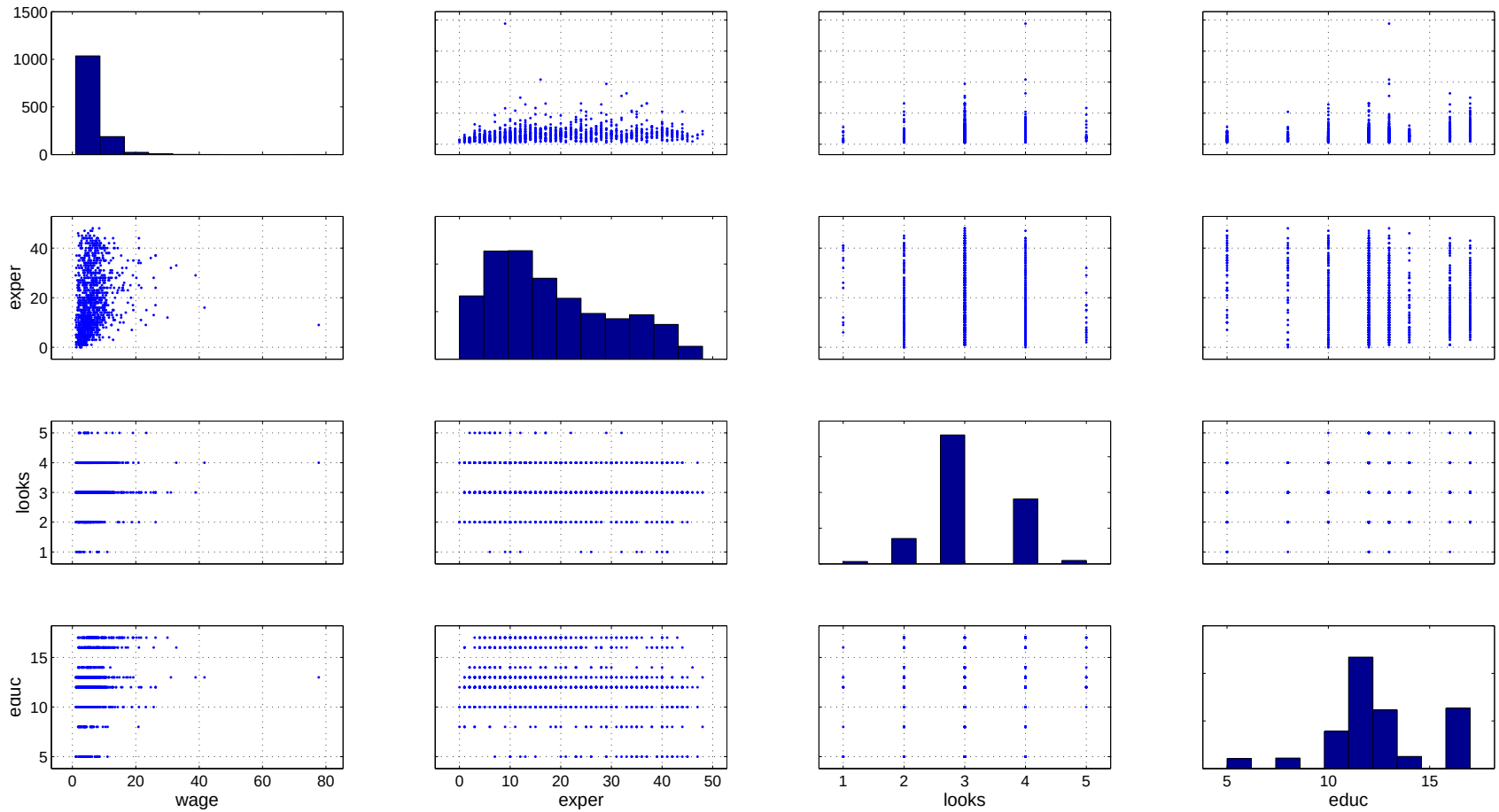
Влияние внешней привлекательности на уровень заработка

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 сфере обслуживания (да/нет).

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

Данные



Данные

В группах $looks = 0$ и $looks = 5$ слишком мало наблюдений.

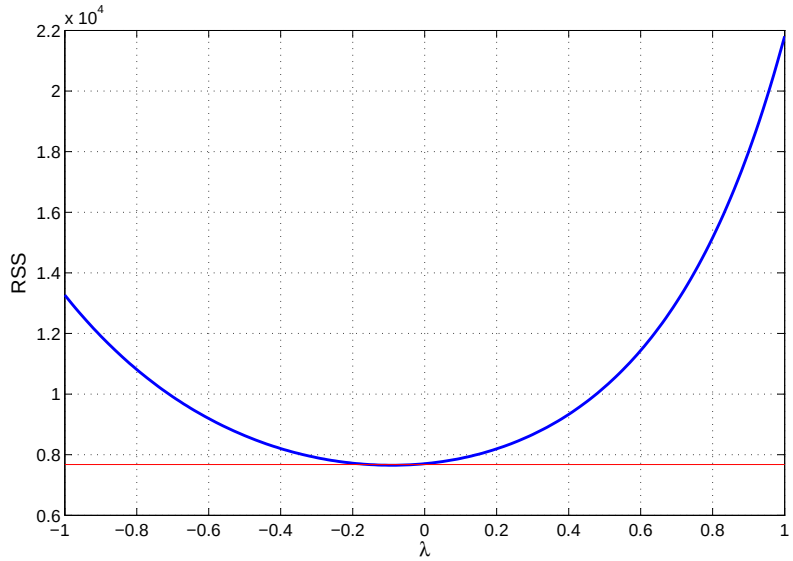
Превратим признак $looks$ в категориальный и закодируем при помощи фиктивных переменных:

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

Преобразование отклика

$$\frac{\max Y}{\min Y} = 76.1961.$$

Найдём преобразование отклика при помощи метода Бокса-Кокса:



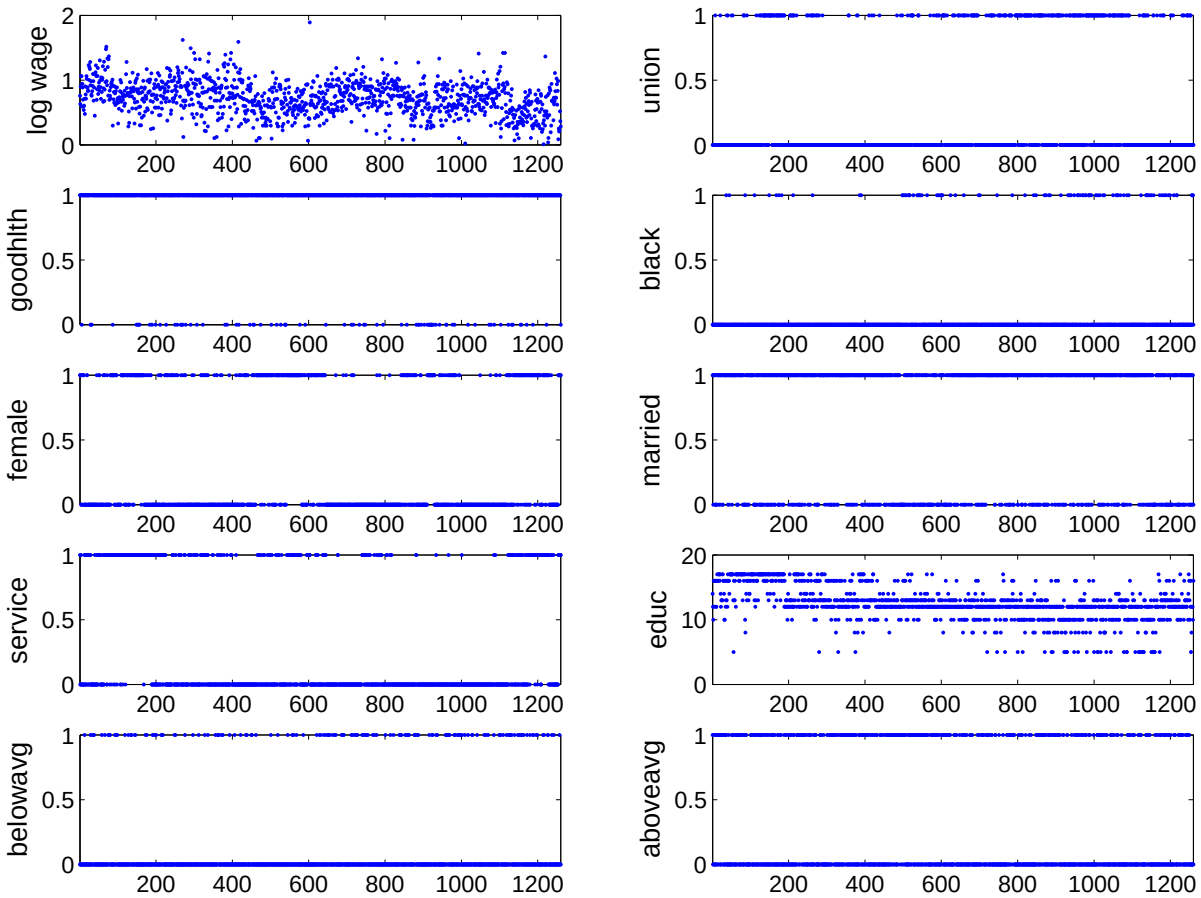
Доверительный интервал для λ определяется как пересечение кривой $RSS(\lambda)$ с линией уровня $\min\limits_{\lambda} RSS\left(\lambda\right) \cdot e^{\chi_1^2(1-\alpha)/n}$.

95% доверительный интервал — $\left(-0.1555, -0.0296\right)$.

Для удобства возьмём всё равно $\lambda = 0$, т. е. будем делать регрессию на логарифм отклика, причём для лучшей интерпретируемости возьмём десятичный логарифм.

Данные

Зависимость от номера наблюдения:



Модель 1

Построим линейную модель без интеракций:

$$\begin{aligned}\log wage = 0.19 + 0.01exper + 0.08union + 0.03goodhlth - 0.02black - \\ - 0.17female + 0.02married - 0.06service + 0.03educ - \\ - 0.06belowavg + 0.002aboveavg.\end{aligned}$$

$F = 74.169, p = 4.7 \times 10^{-119}, R^2 = 0.373, R_a^2 = 0.367.$

Критерий	p-value
Шапиро-Уилка (нормальность)	1.4×10^{-10}
знаковых рангов (несмещённость)	0.8383
Бройша-Пагана (гомоскедастичность)	4.9×10^{-6}
Дарбина-Уотсона (некоррелированность)	0.0032

Признаки, коэффициенты при которых значимо отличаются от нуля согласно критерию Стьюдента: $exper$, $union$, $female$, $service$, $educ$, $belowavg$.

Модель 2

Редуцированная модель:

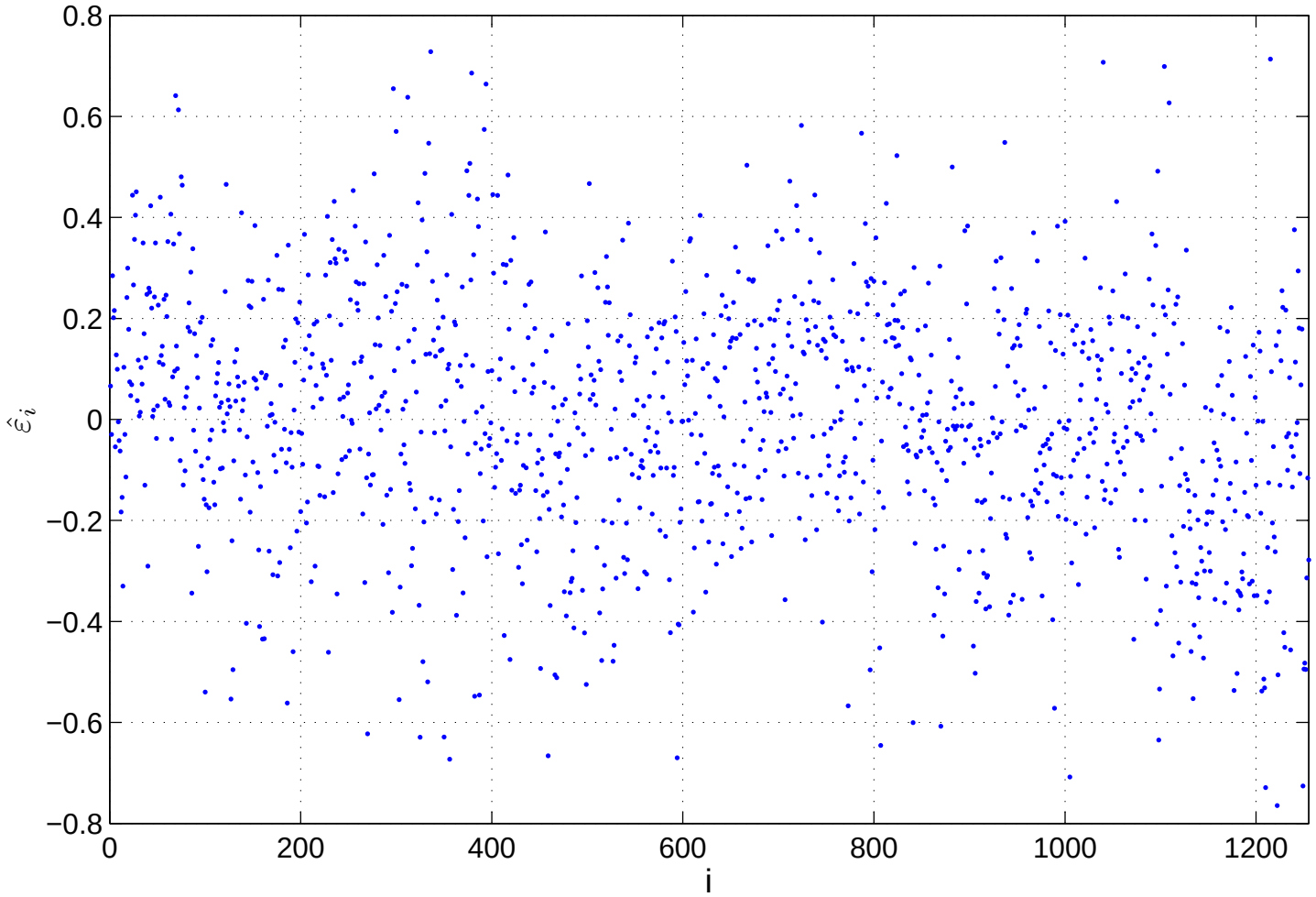
$$\log wage = 0.23 + 0.01exper + 0.08union - 0.17female - 0.06service + \\ + 0.04educ - 0.05belowavg + 0.002aboveavg.$$

$F = 104.586, p = 1.7 \times 10^{-120}, R^2 = 0.369, R_a^2 = 0.366.$

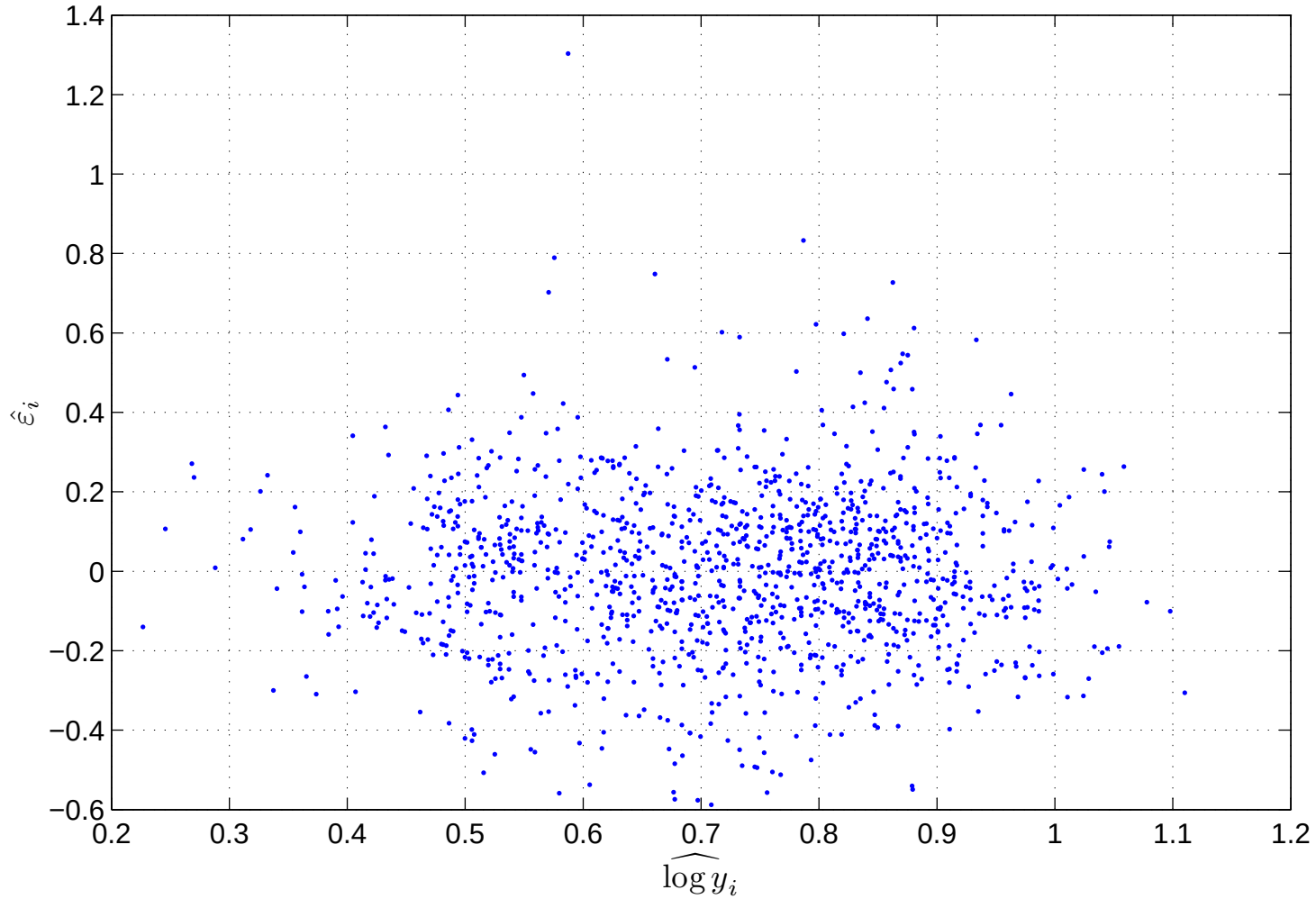
Критерий	p-value
Шапиро-Уилка (нормальность)	2.7×10^{-10}
знаковых рангов (несмещённость)	0.8851
Бройша-Пагана (гомоскедастичность)	5.7×10^{-6}
Дарбина-Уотсона (некоррелированность)	0.0035

Значимы все признаки, кроме $aboveavg$.

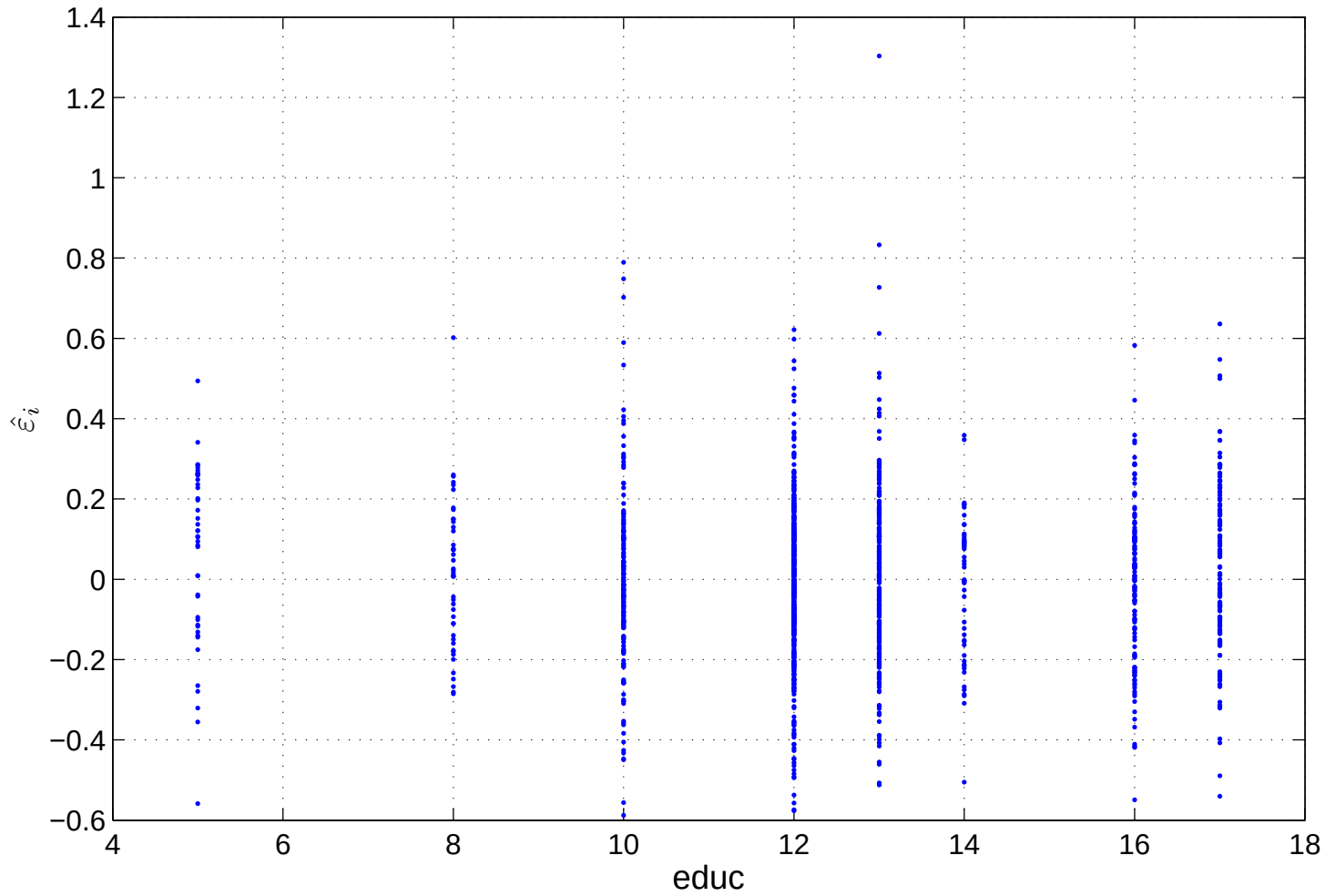
Остатки модели 2



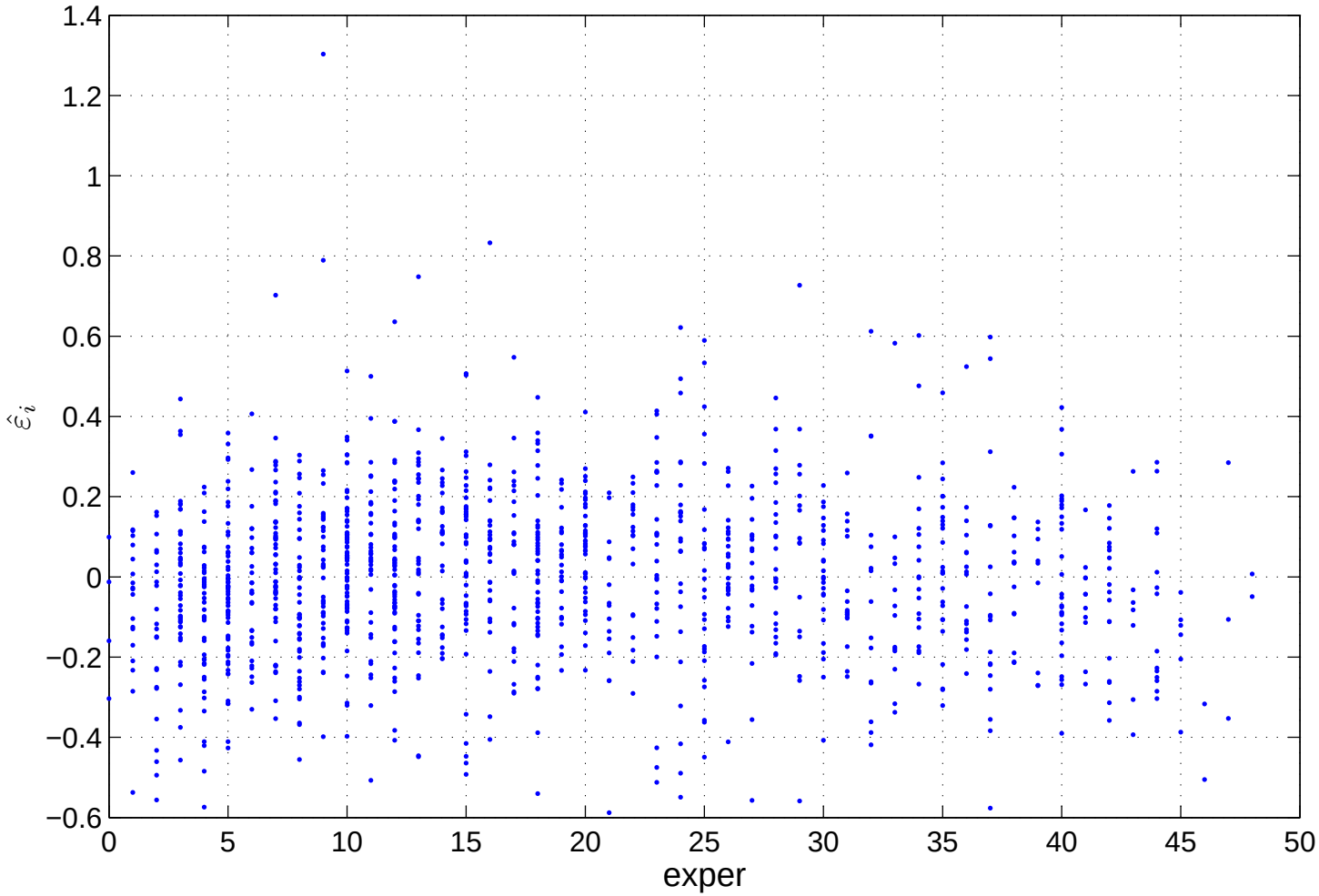
Остатки модели 2



Остатки модели 2



Остатки модели 2



Модель 3

Модель с квадратом признака $exper$:

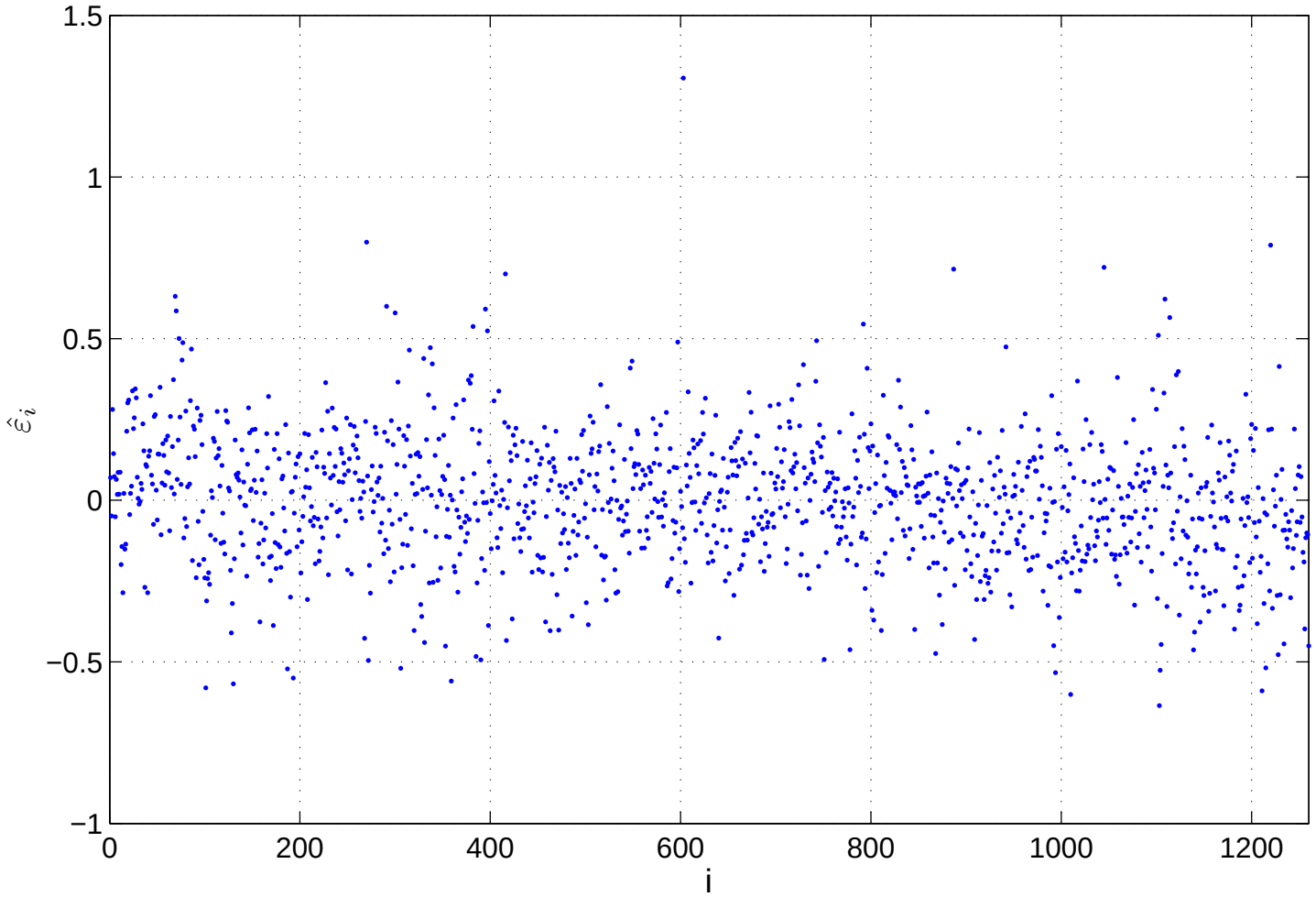
$$\log wage = 0.17 + 0.02exper - 0.0003exper^2 + 0.08union - 0.17female - \\ - 0.07service + 0.03educ - 0.06belowavg + 0.003aboveavg.$$

$F = 99.878, p = 1.7 \times 10^{-128}, R^2 = 0.390, R_a^2 = 0.386.$

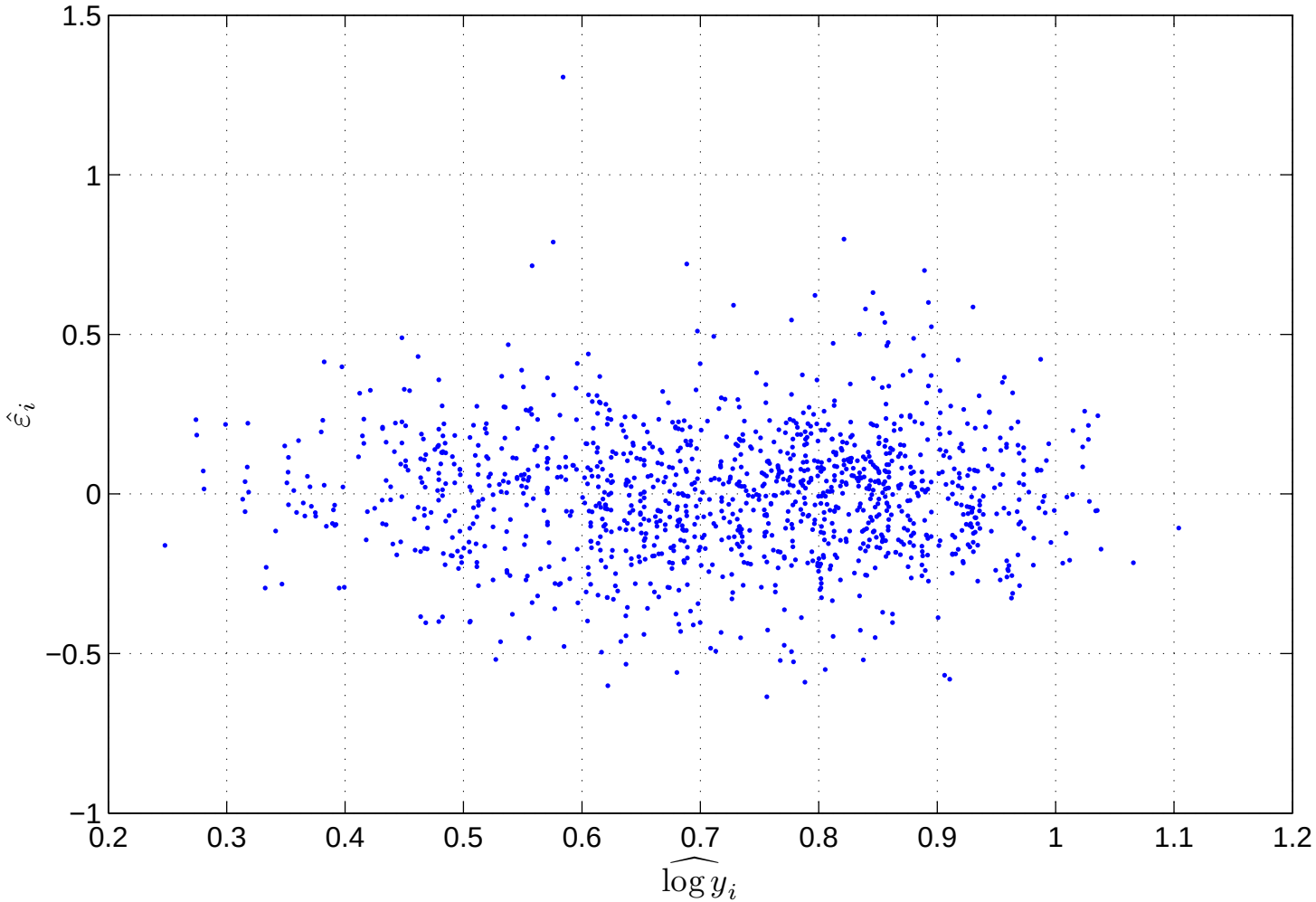
Критерий	p-value
Шапиро-Уилка (нормальность)	8.02×10^{-11}
знаковых рангов (несмещённость)	0.8803
Бройша-Пагана (гомоскедастичность)	4.7×10^{-5}
Дарбина-Уотсона (некоррелированность)	0.0029

Значимы все признаки, кроме $aboveavg$.

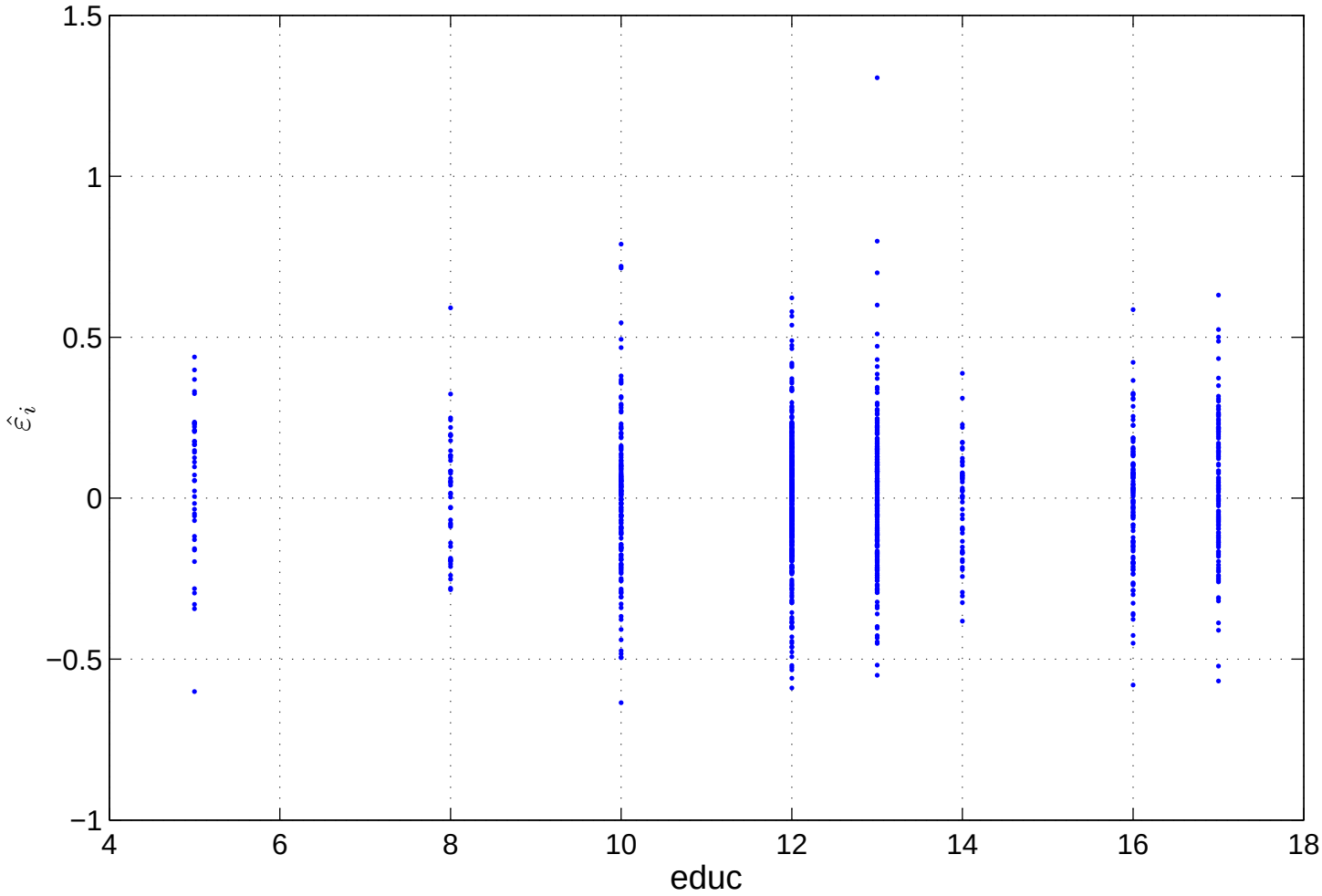
Остатки модели 3



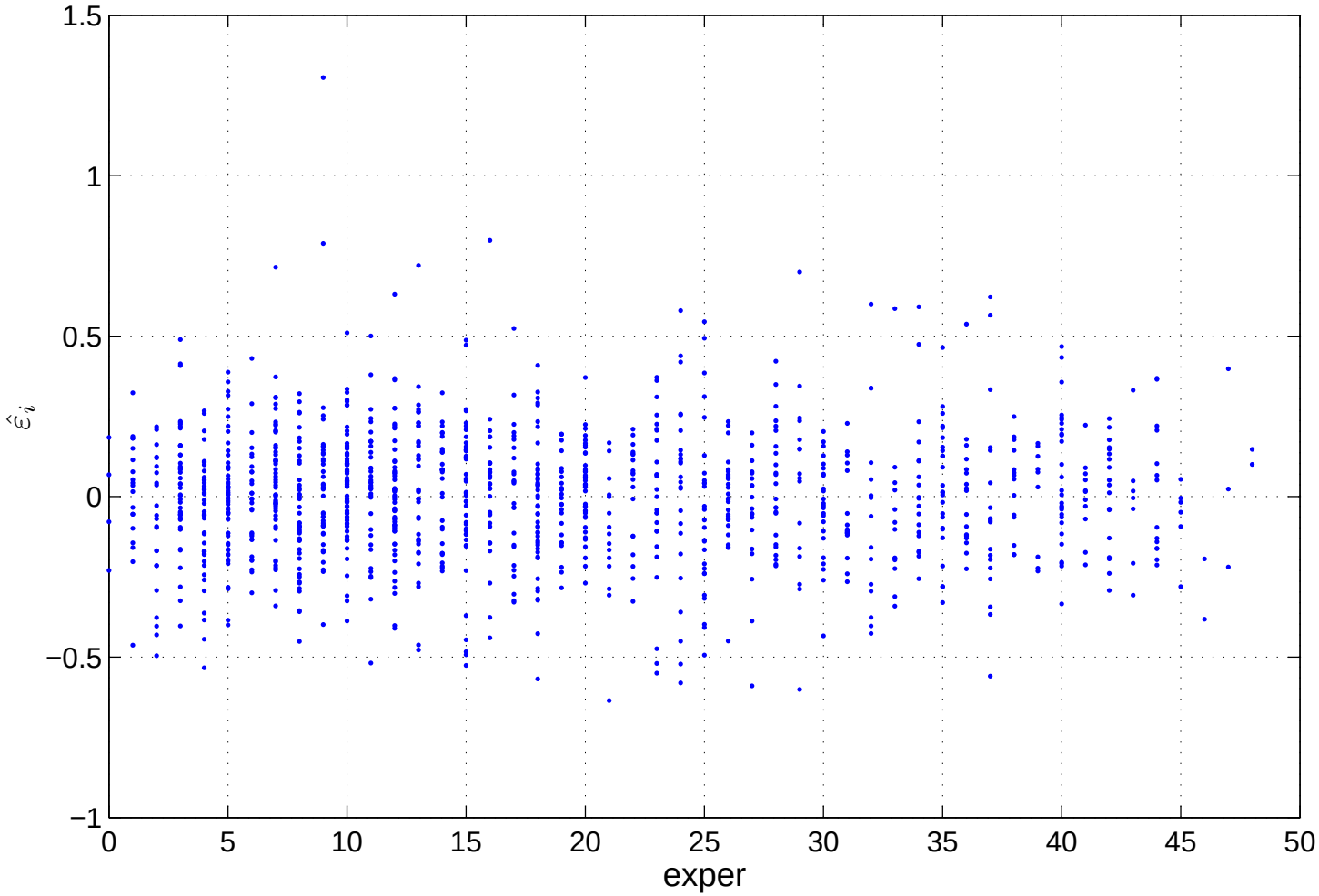
Остатки модели 3



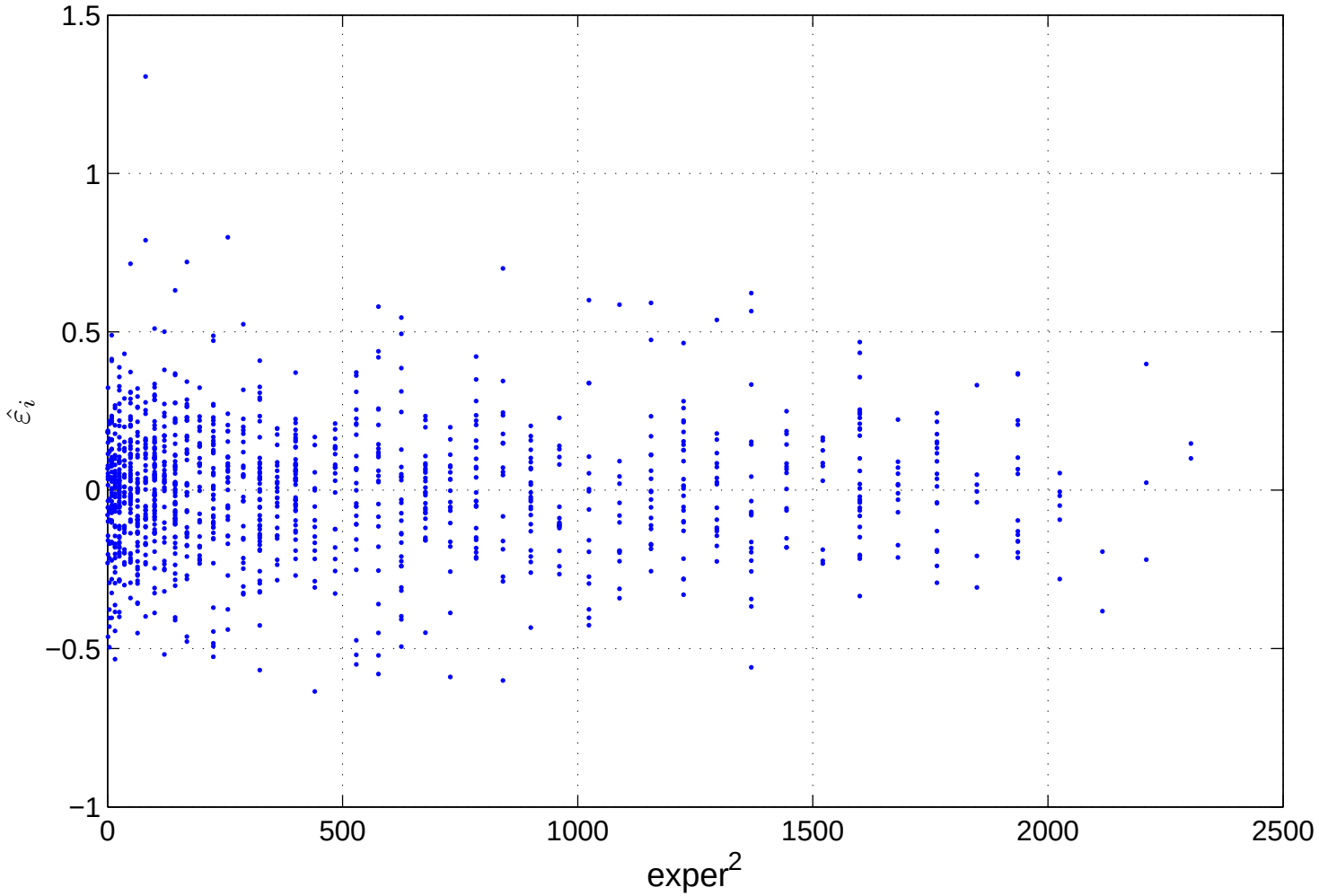
Остатки модели 3



Остатки модели 3



Остатки модели 3



Модель 4

Сделаем пошаговую регрессию со всеми попарными взаимодействиями (кроме взаимодействия фиктивных переменных) и квадратами числовых признаков, затем добавим к отобранным признакам все «чистые» признаки, входящие в значимые интеракции.

$$\begin{aligned}\log wage = {} & 0.3 + 0.02exper + 0.12union + 0.02goodhlth - 0.03black - \\ & - 0.16female + 0.03married - 0.04service - 0.002educ + \\ & + 0.02belowavg - 0.0001aboveavg - 0.002exper * union - \\ & - 0.003exper * female - 0.002exper * service + 0.12union * black - \\ & - 0.15goodhlth * black + 0.08goodhlth * female - \\ & - 0.09goodhlth * belowavg + 0.21black * female - 0.06female * married - \\ & - 0.0003exper^2 + 0.001 * educ^2.\end{aligned}$$

$F = 43.04, p = 3.9 \times 10^{-131}, R^2 = 0.422, R_a^2 = 0.412.$

Критерий	p-value
Шапиро-Уилка (нормальность)	1.02×10^{-9}
знаковых рангов (несмещённость)	0.8592
Бройша-Пагана (гомоскедастичность)	3×10^{-5}
Дарбина-Уотсона (некоррелированность)	1.34×10^{-4}

Модель 5

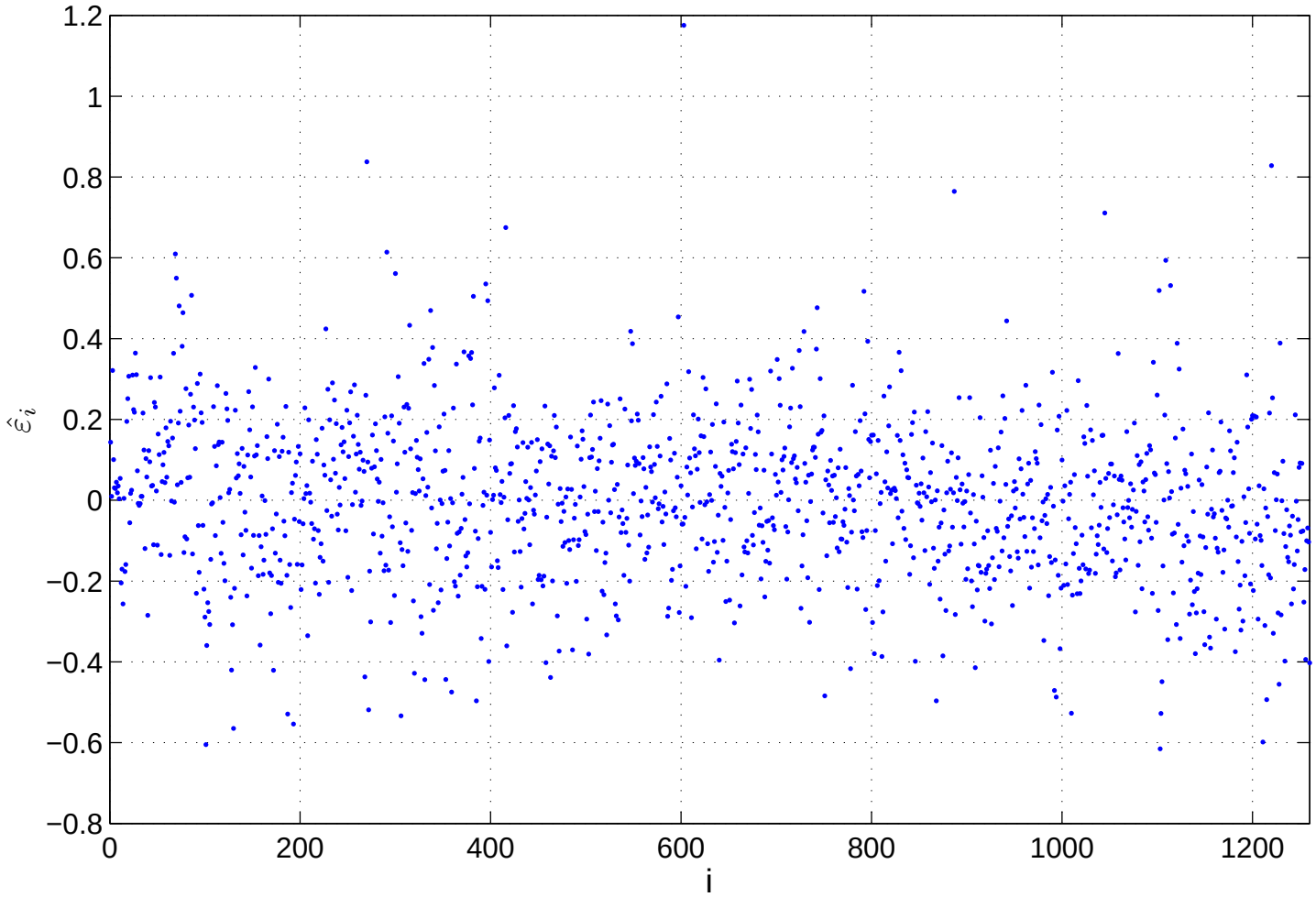
Чтобы упростить модель и повысить её интерпретируемость, исключим три взаимодействия, для которых критерий Стьюдента даёт достигаемый уровень значимости меньше 0.05:

$$\begin{aligned}\log wage = {} & 0.3 + 0.02exper + 0.12union + 0.04goodhlth - 0.05black - \\ & - 0.08female + 0.03married - 0.08service - 0.004educ - \\ & - 0.06belowavg + 0.0002aboveavg - 0.002exper * union - \\ & - 0.004exper * female + 0.11union * black - 0.13goodhlth * black + \\ & + 0.2black * female - 0.06female * married - 0.0003exper^2 + \\ & + 0.002 * educ^2.\end{aligned}$$

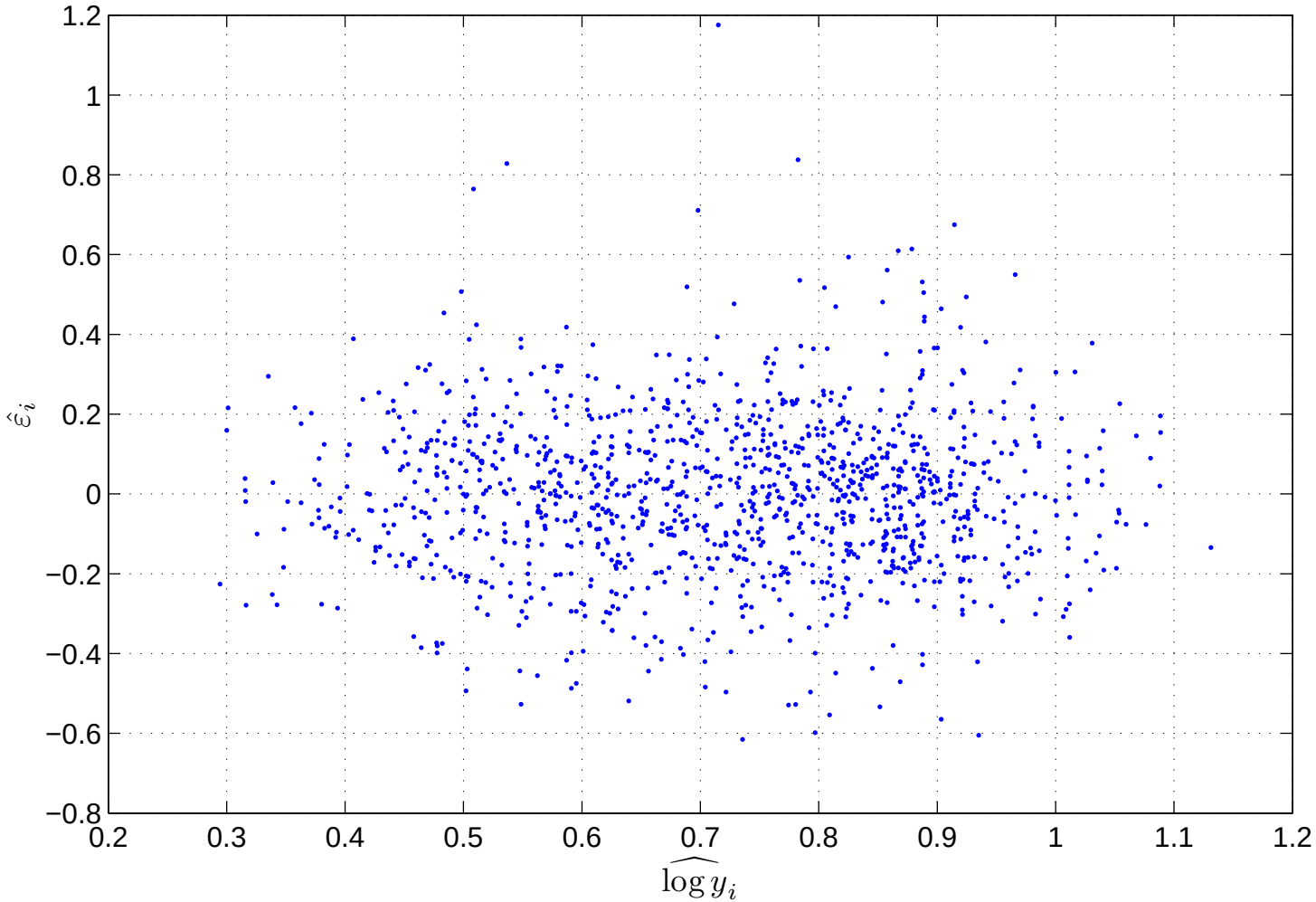
$F = 49.6, p = 4.9 \times 10^{-132}, R^2 = 0.418, R_a^2 = 0.41.$

Критерий	p-value
Шапиро-Уилка (нормальность)	4×10^{-10}
знаковых рангов (несмещённость)	0.8803
Бройша-Пагана (гомоскедастичность)	1.9×10^{-5}
Вулдриджа (ослабление гомоскедастичности)	0.2841
Дарбина-Уотсона (некоррелированность)	2.14×10^{-4}

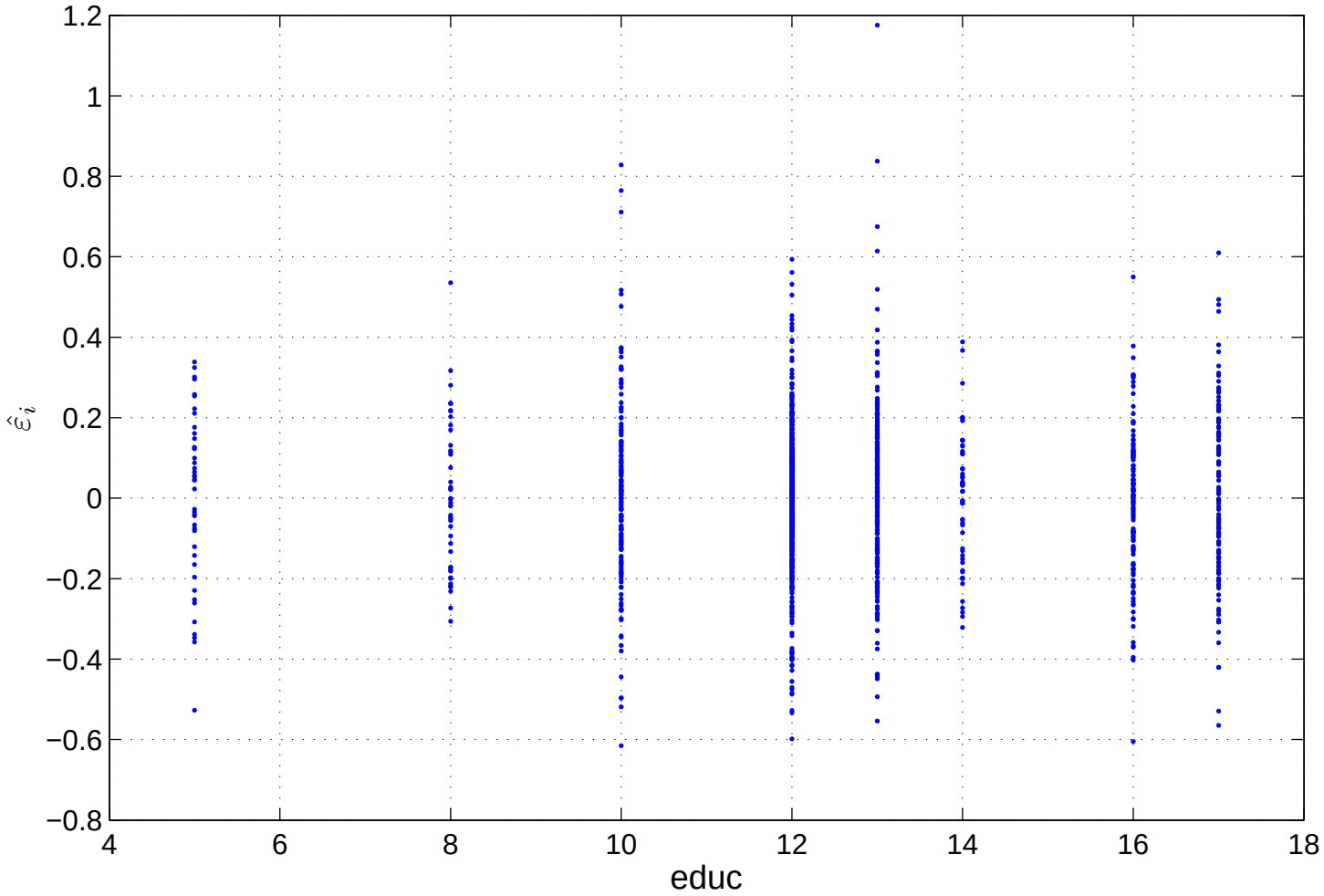
Остатки модели 5



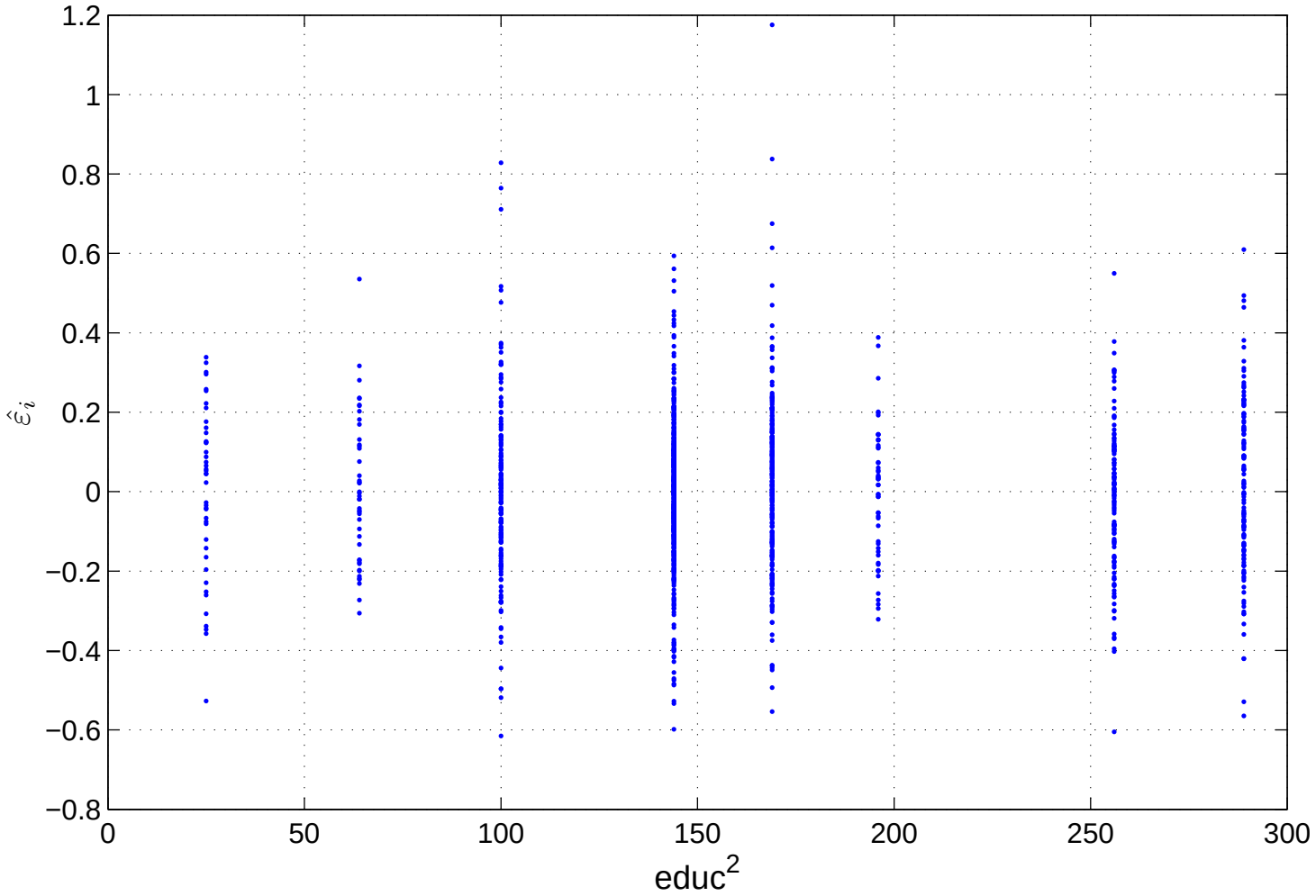
Остатки модели 5



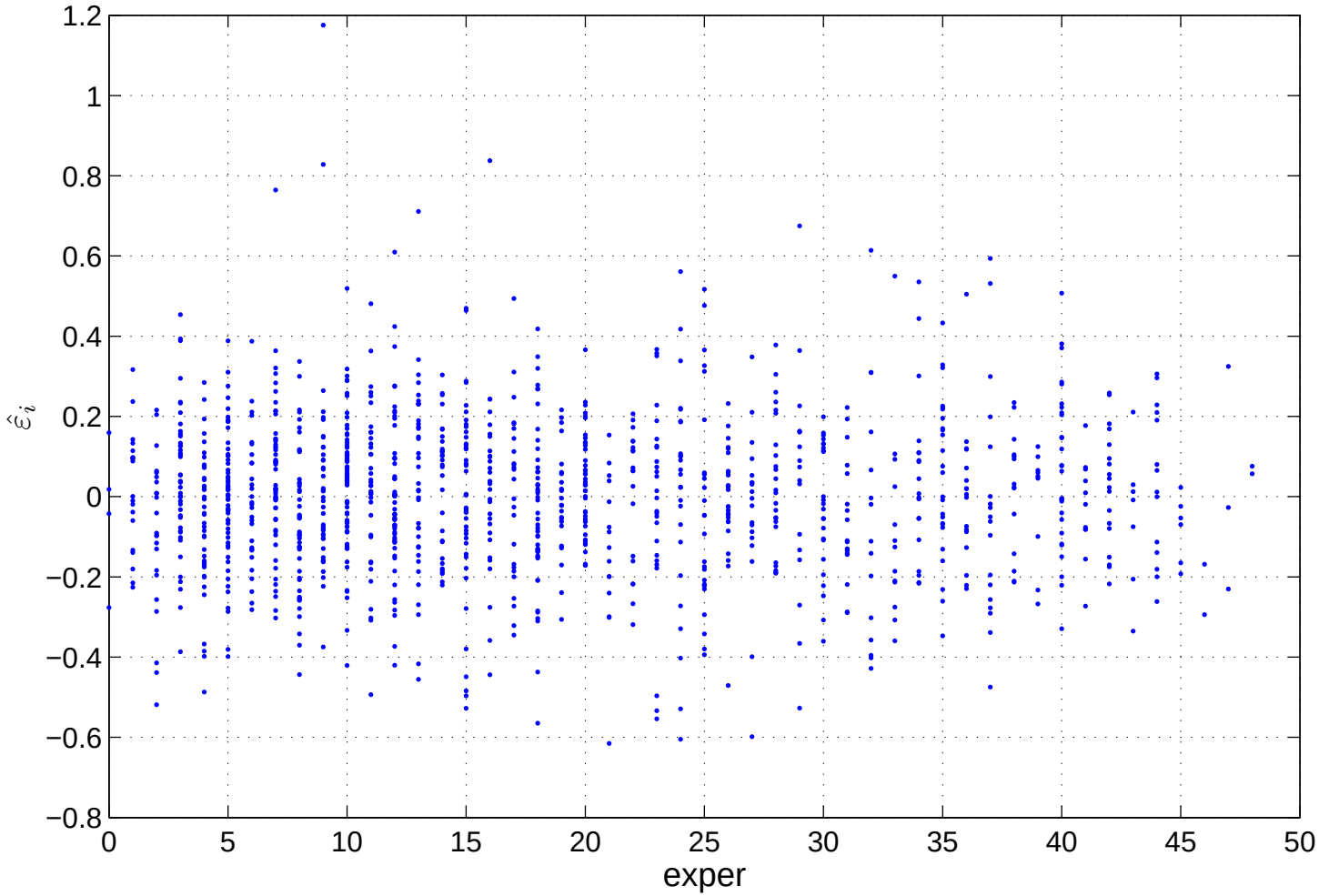
Остатки модели 5



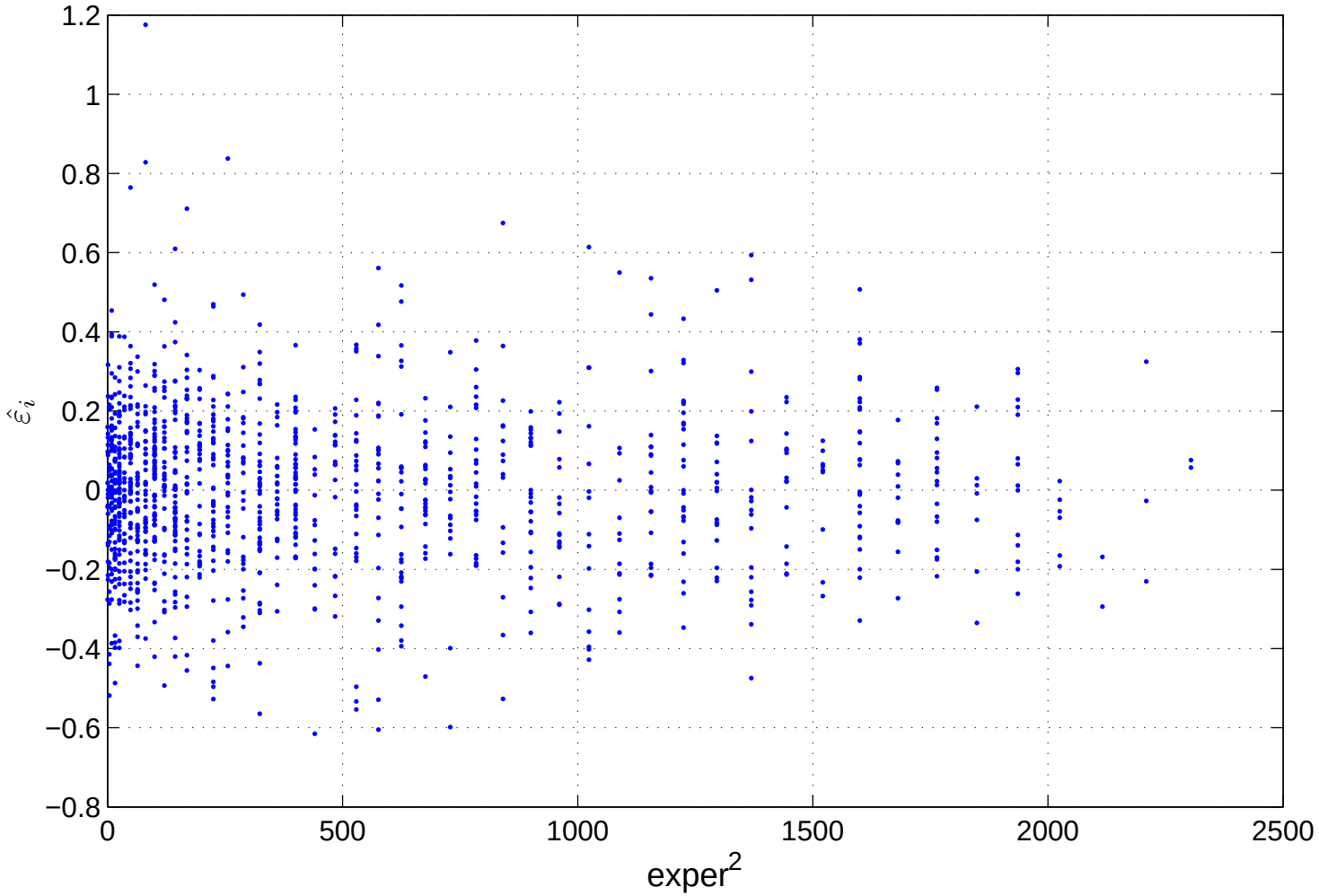
Остатки модели 5



Остатки модели 5

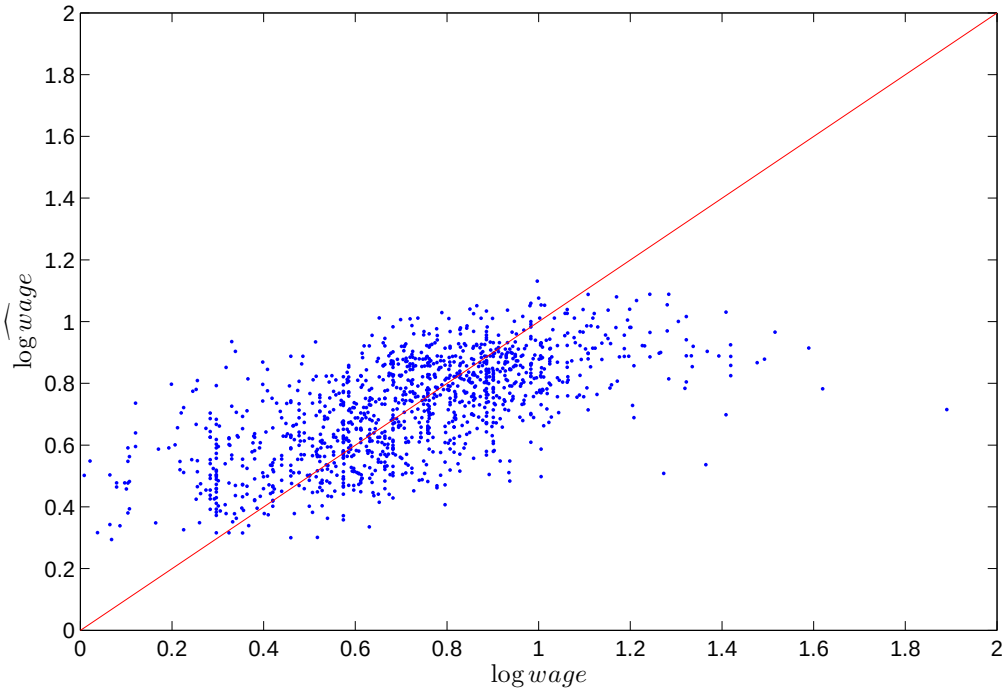


Остатки модели 5



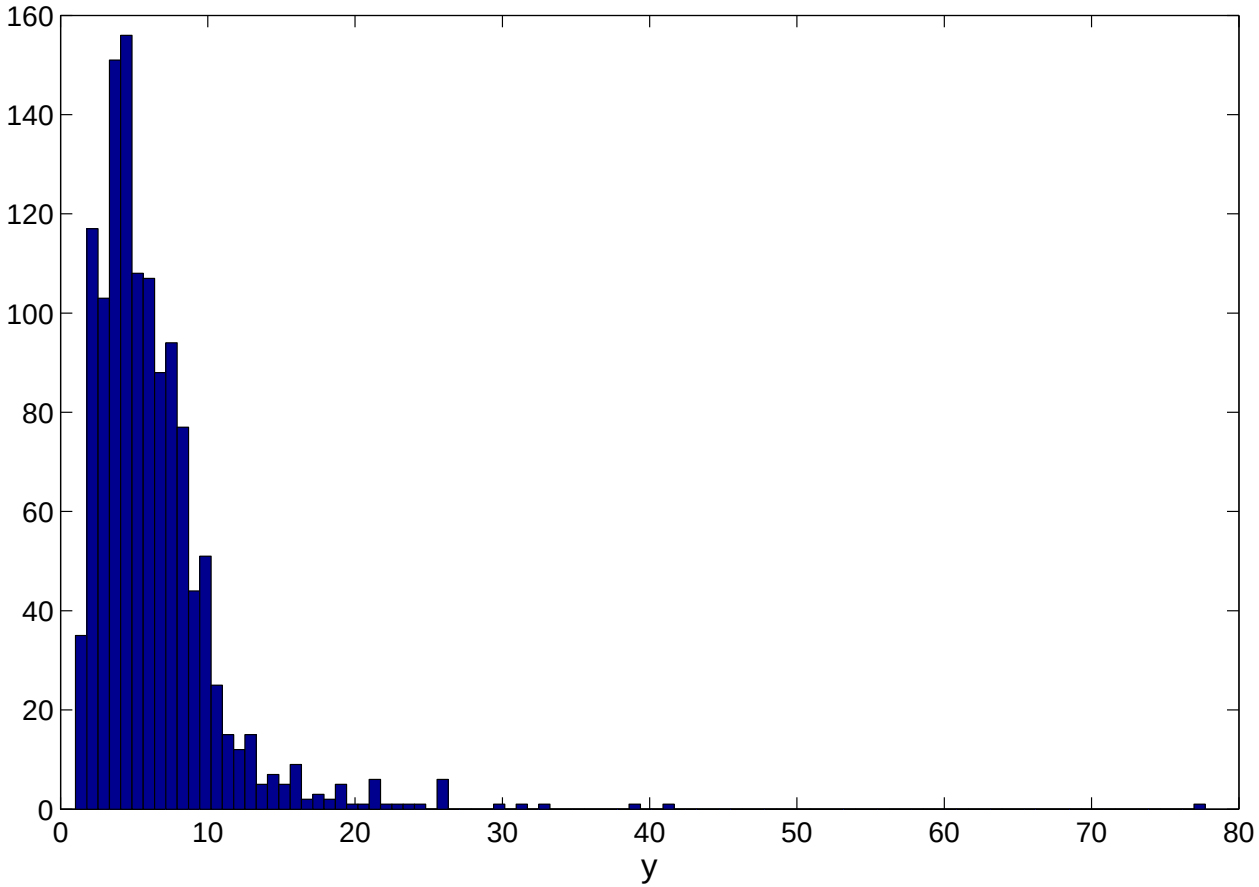
Результат

Итоговая модель объясняет 42% вариации логарифма отклика:



С учётом дополнительных факторов, участники опроса с привлекательностью ниже среднего получают на 6% больше (95% доверительный интервал $(2.7\%, 9.71\%)$), а с привлекательностью выше среднего — на 0.02% меньше (95% доверительный интервал $(-2.50\%, 2.53\%)$).

Выбросы

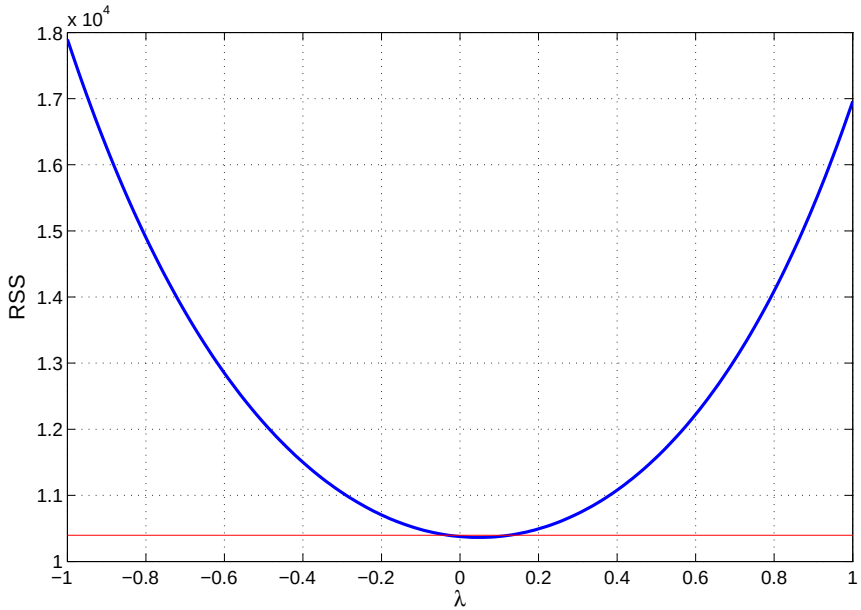


Больше 30 долларов в час в выборке получает только 5 человек. Исключим их.

Преобразование отклика

$$\frac{\max Y}{\min Y} = 29.3922.$$

Найдём преобразование отклика при помощи метода Бокса-Кокса:



Доверительный интервал для λ определяется как пересечение кривой $RSS(\lambda)$ с линией уровня $\min_\lambda RSS\left(\lambda\right) \cdot e^{\chi_1^2(1-\alpha)/n}$.
95% доверительный интервал — $\left(-0.028, 0.124\right)$.
Возьмём $\lambda = 0$, т. е. будем делать регрессию на логарифм отклика, причём для лучшей интерпретируемости возьмём десятичный логарифм.

Модель 1

Построим линейную модель без интеракций:

$$\log wage = 0.43 + 0.002exper - 0.01union + 0.01goodhlth - 0.03black - \\ - 0.08female + 0.02married - 0.03service + 0.02educ + \\ + 0.01belowavg - -0.01aboveavg.$$

$F = 13.18, p = 3.6 \times 10^{-22}, R^2 = 0.0958, R_a^2 = 0.0855.$

Критерий	p-value
Шапиро-Уилка (нормальность)	0.0067
знаковых рангов (несмещённость)	0.5081
Бройша-Пагана (гомоскедастичность)	0.0062
Дарбина-Уотсона (некоррелированность)	4.1×10^{-10}

Признаки, коэффициенты при которых значимо отличаются от нуля согласно критерию Стьюдента: $exper$, $female$, $educ$.

Модель 2

Редуцированная модель:

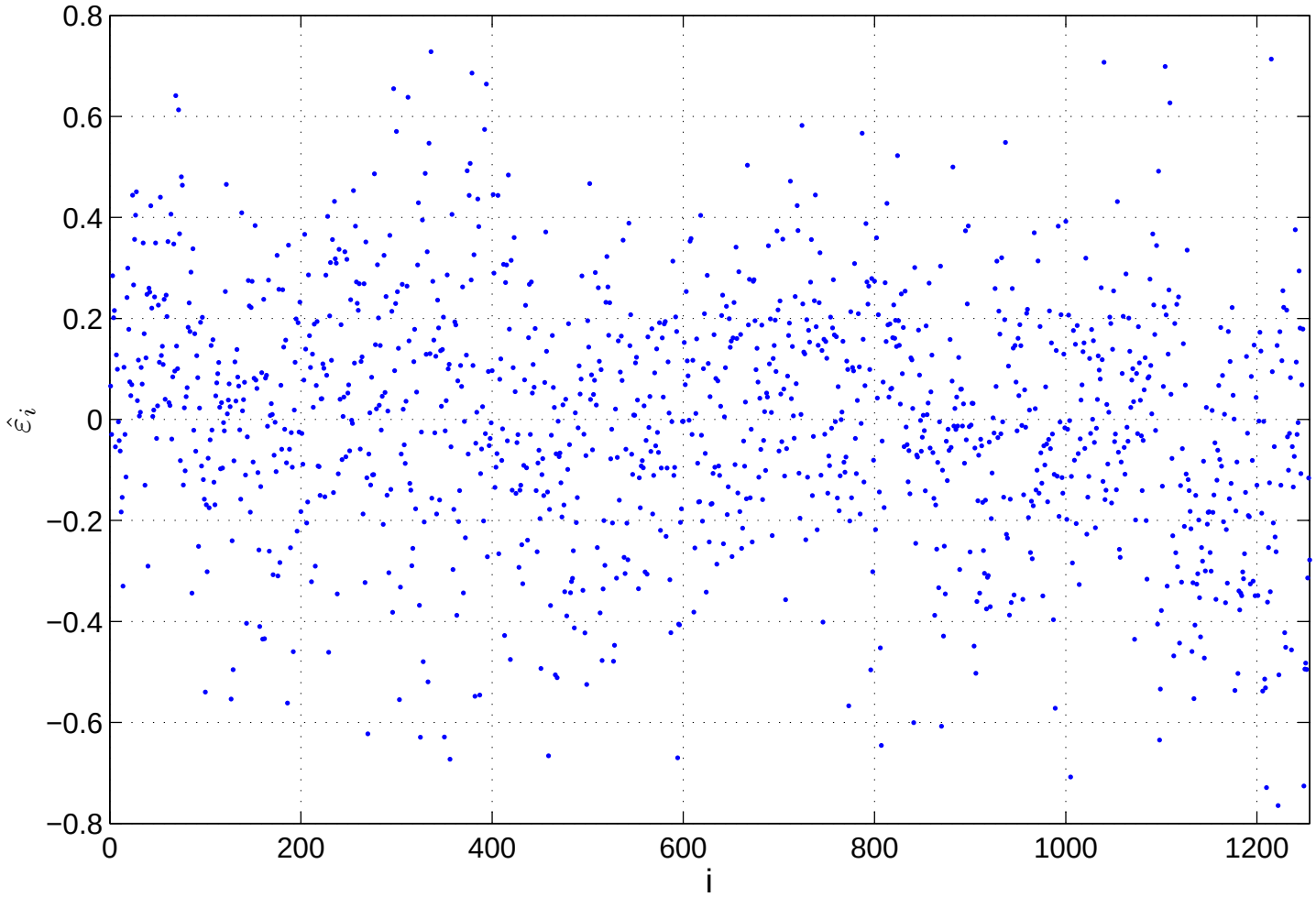
$$\log wage = 0.45 + 0.002exper - 0.1female + 0.02educ -$$
$$+ 0.01belowavg - 0.01aboveavg.$$

$F = 24.88, p = 5.8 \times 10^{-24}, R^2 = 0.0906, R_a^2 = 0.0869.$

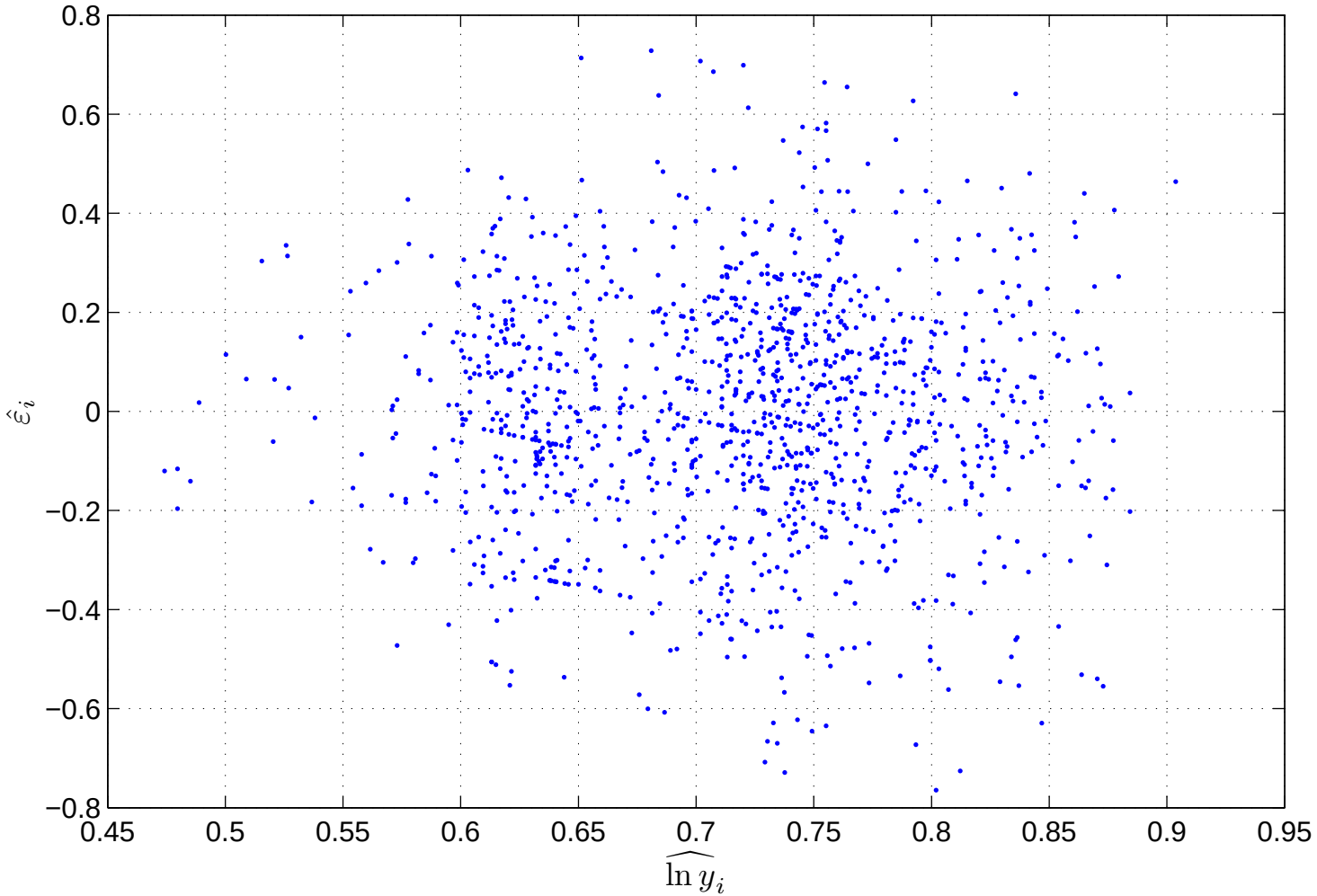
Критерий	p-value
Шапиро-Уилка (нормальность)	0.0032
знаковых рангов (несмещённость)	0.4832
Бройша-Пагана (гомоскедастичность)	0.0045
Дарбина-Уотсона (некоррелированность)	1×10^{-9}

Значимы все признаки, кроме $belowavg$ и $aboveavg$.

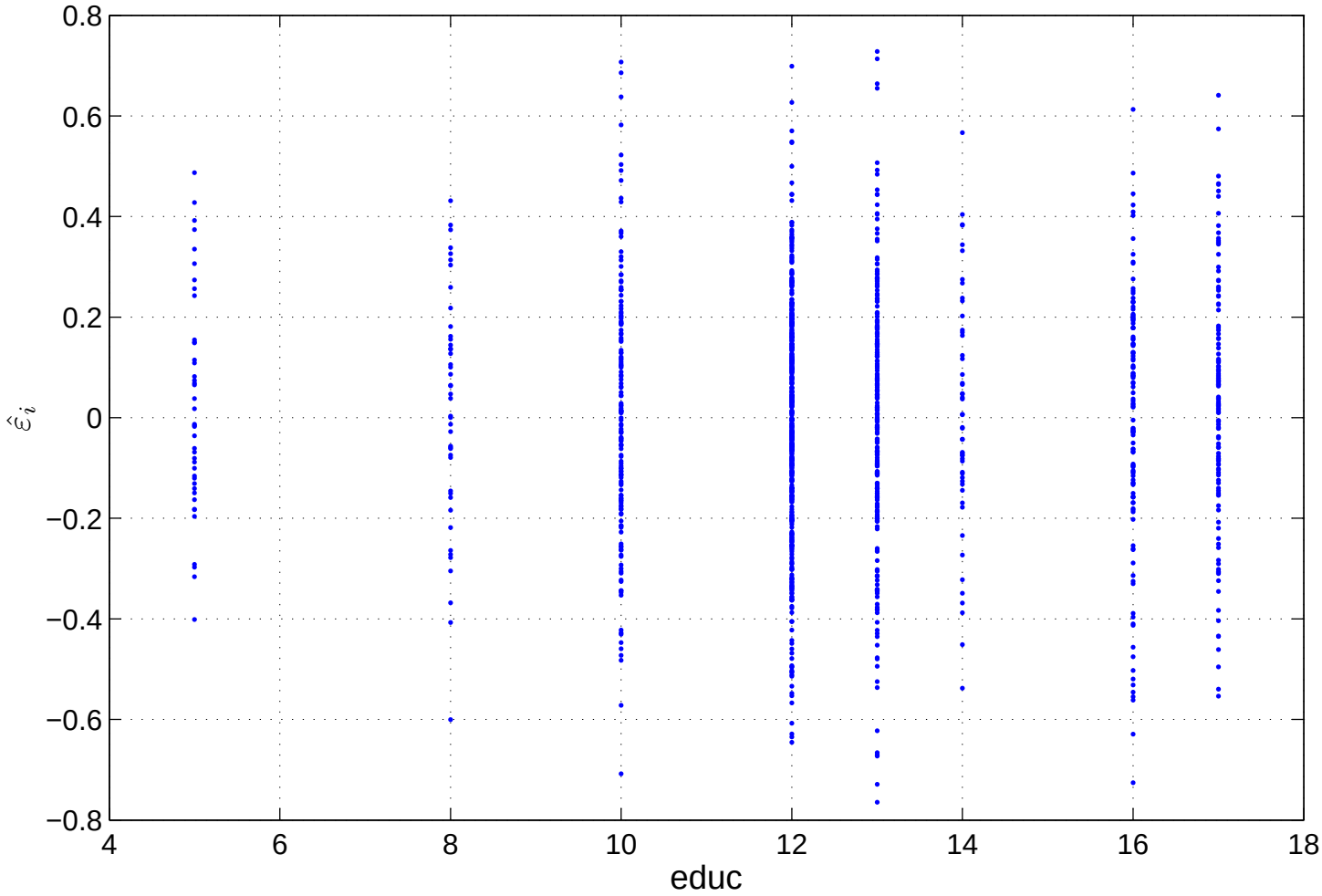
Остатки модели 2



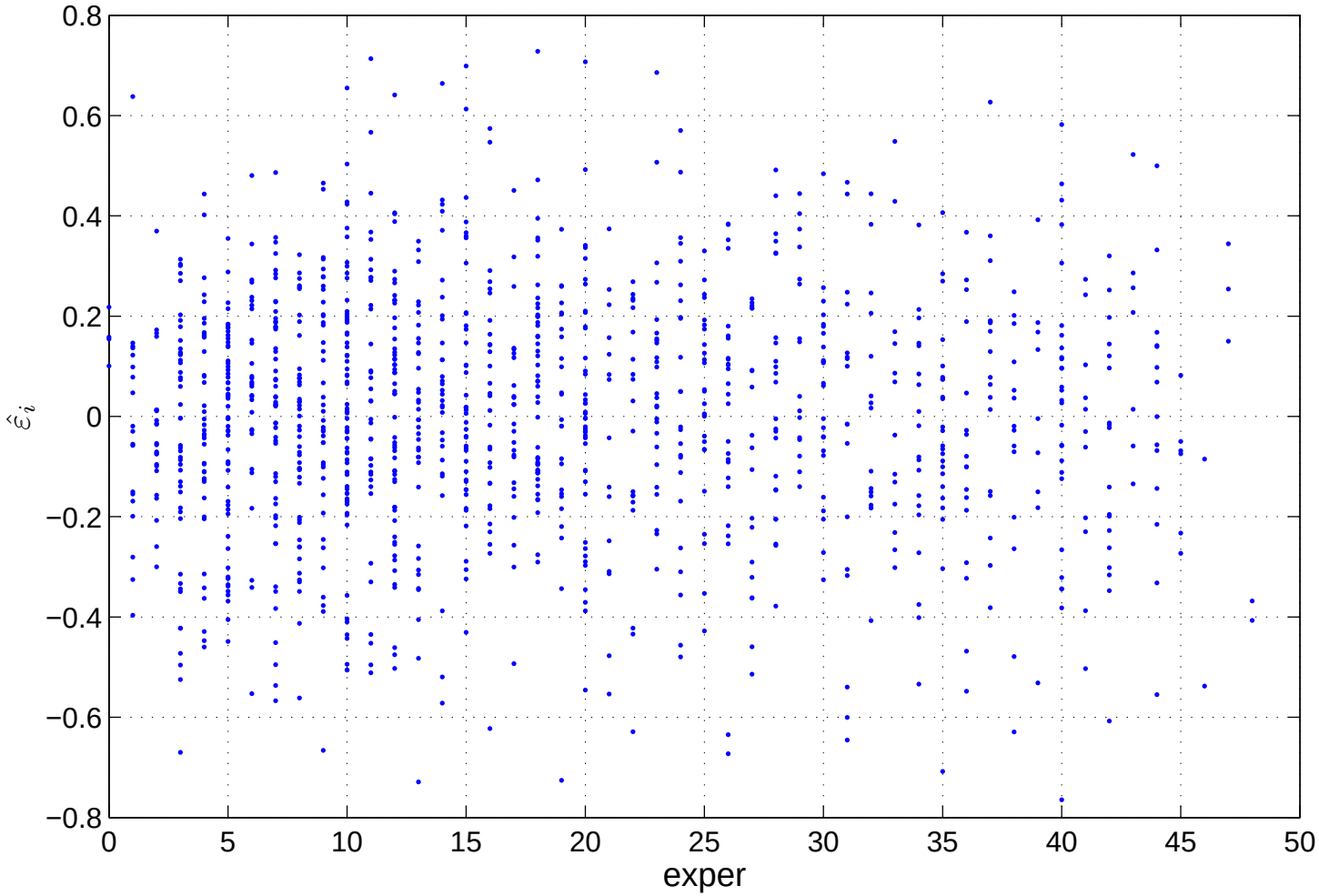
Остатки модели 2



Остатки модели 2



Остатки модели 2



Модель 3

Сделаем пошаговую регрессию со всеми попарными взаимодействиями (кроме взаимодействия фиктивных переменных) и квадратами числовых признаков, затем добавим к отобранным признакам все «чистые» признаки, входящие в значимые интеракции.

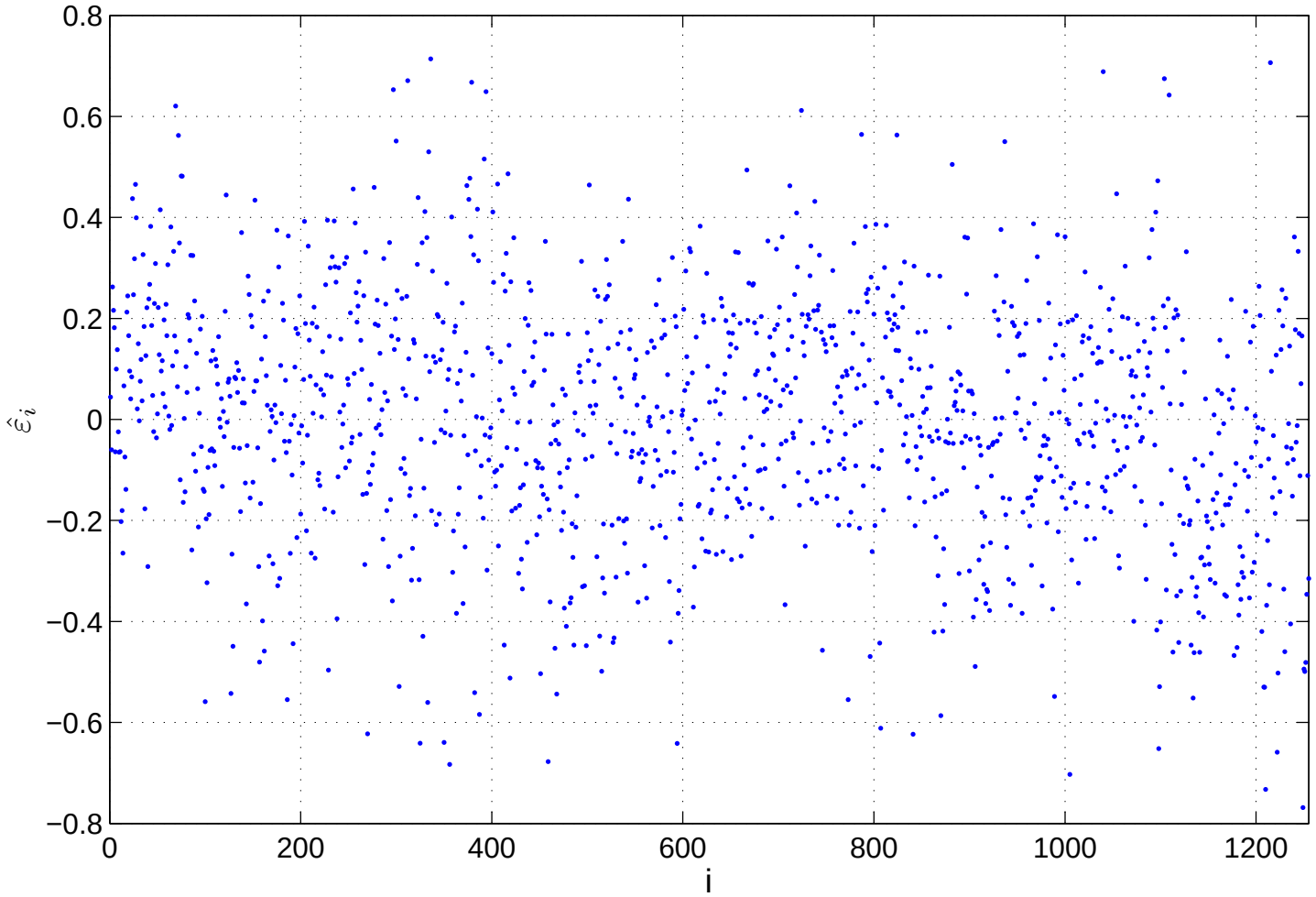
$$\begin{aligned}\ln wage = 0.55 &+ 0.007exper - 0.006union + 0.03goodhlth - 0.08female + \\ &+ 0.003married - 0.003service - 0.006educ + 0.01belowavg - \\ &- 0.007aboveavg - 0.04union * aboveavg - 0.08goodhlth * service + \\ &+ 0.06married * service - 0.0001exper^2 + 0.001 * educ^2.\end{aligned}$$

$F = 10.5, p = 8.7 \times 10^{-23}, R^2 = 0.1060, R_a^2 = 0.0959.$

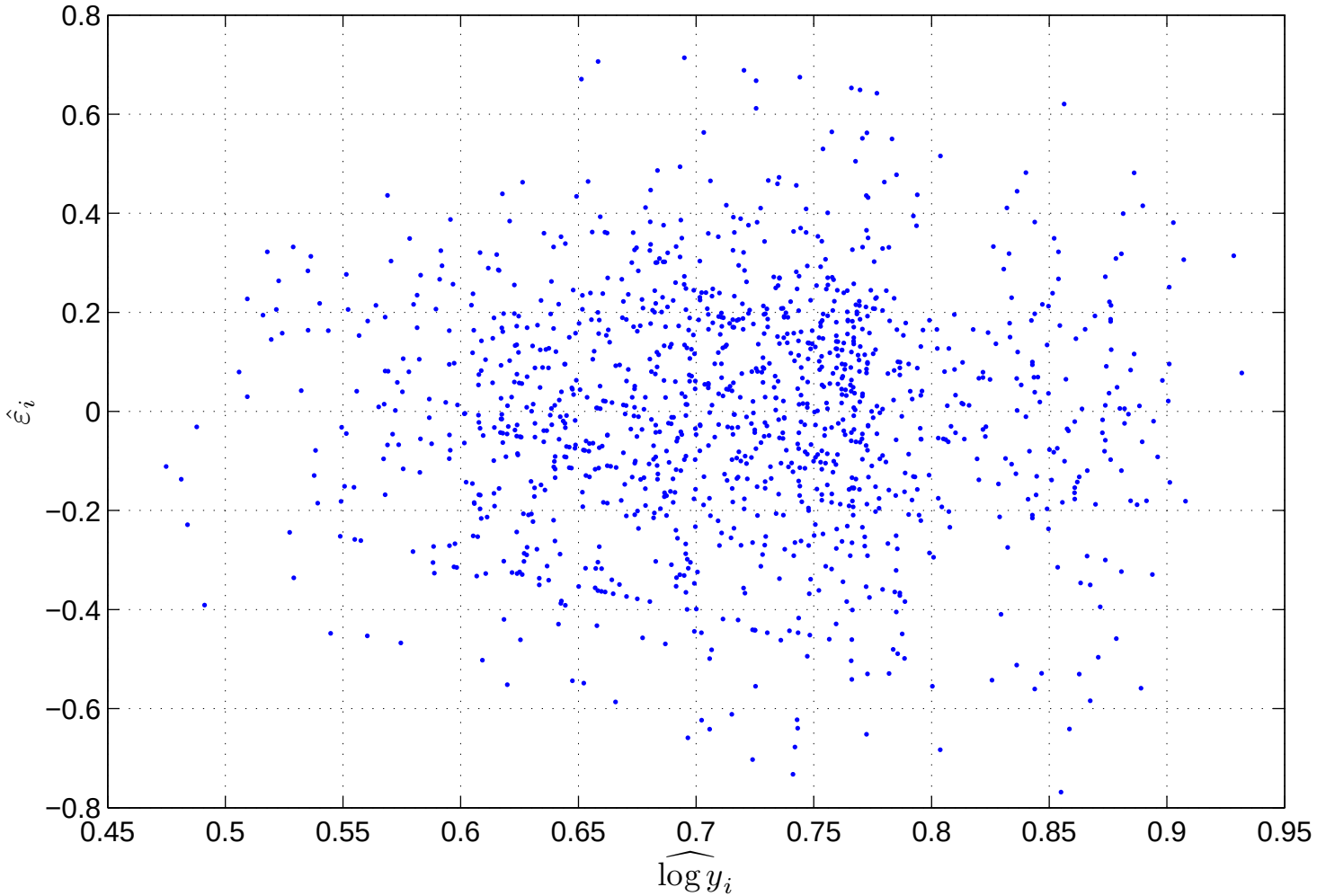
Критерий	p-value
Шапиро-Уилка (нормальность)	0.0038
знаковых рангов (несмещённость)	0.4783
Бройша-Пагана (гомоскедастичность)	0.0031
Вулдриджа (ослабление гомоскедастичности)	0.0484
Дарбина-Уотсона (некоррелированность)	3.05×10^{-10}

По критерию Стьюдента значимы: $exper$, $female$, $exper^2$.

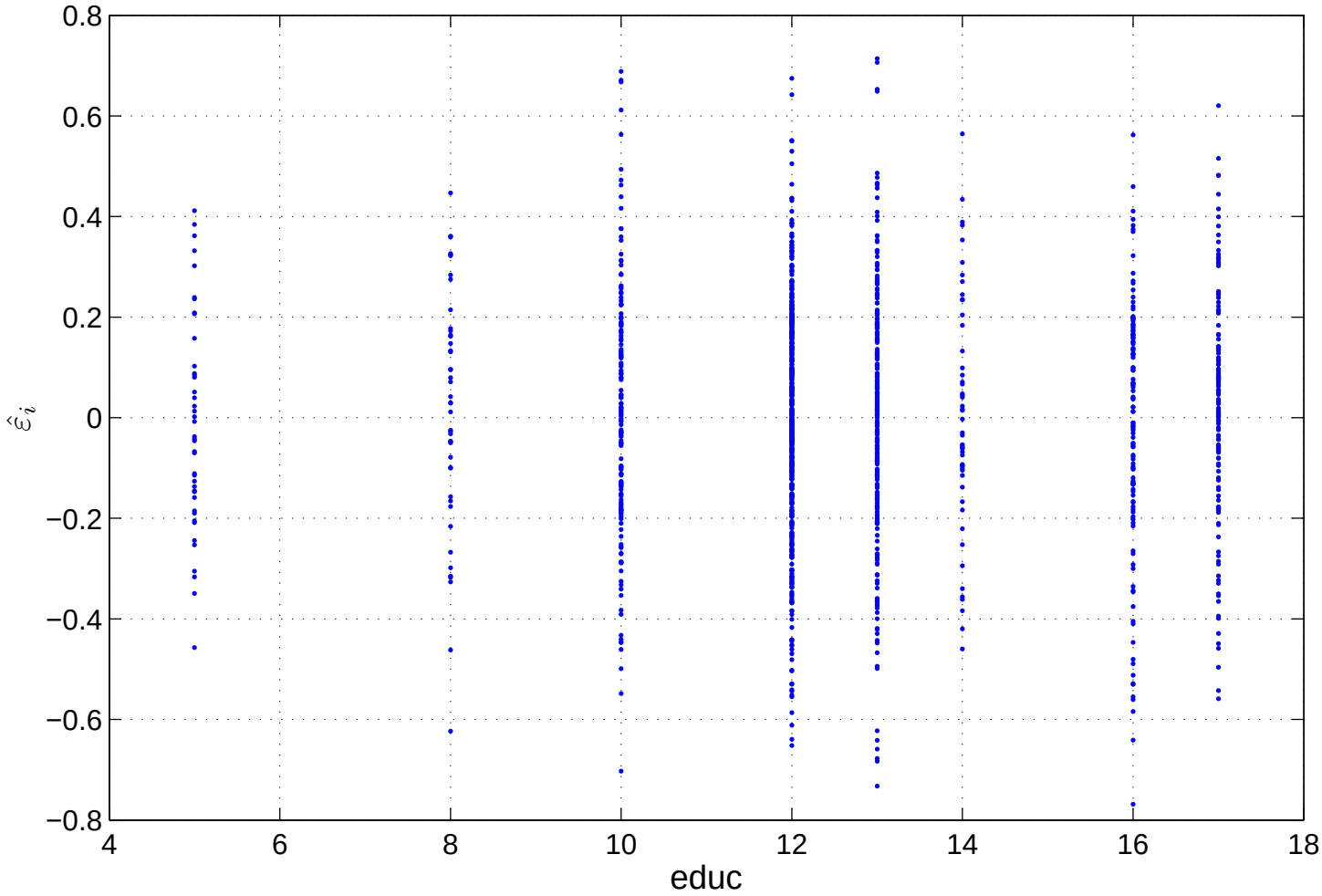
Остатки модели 3



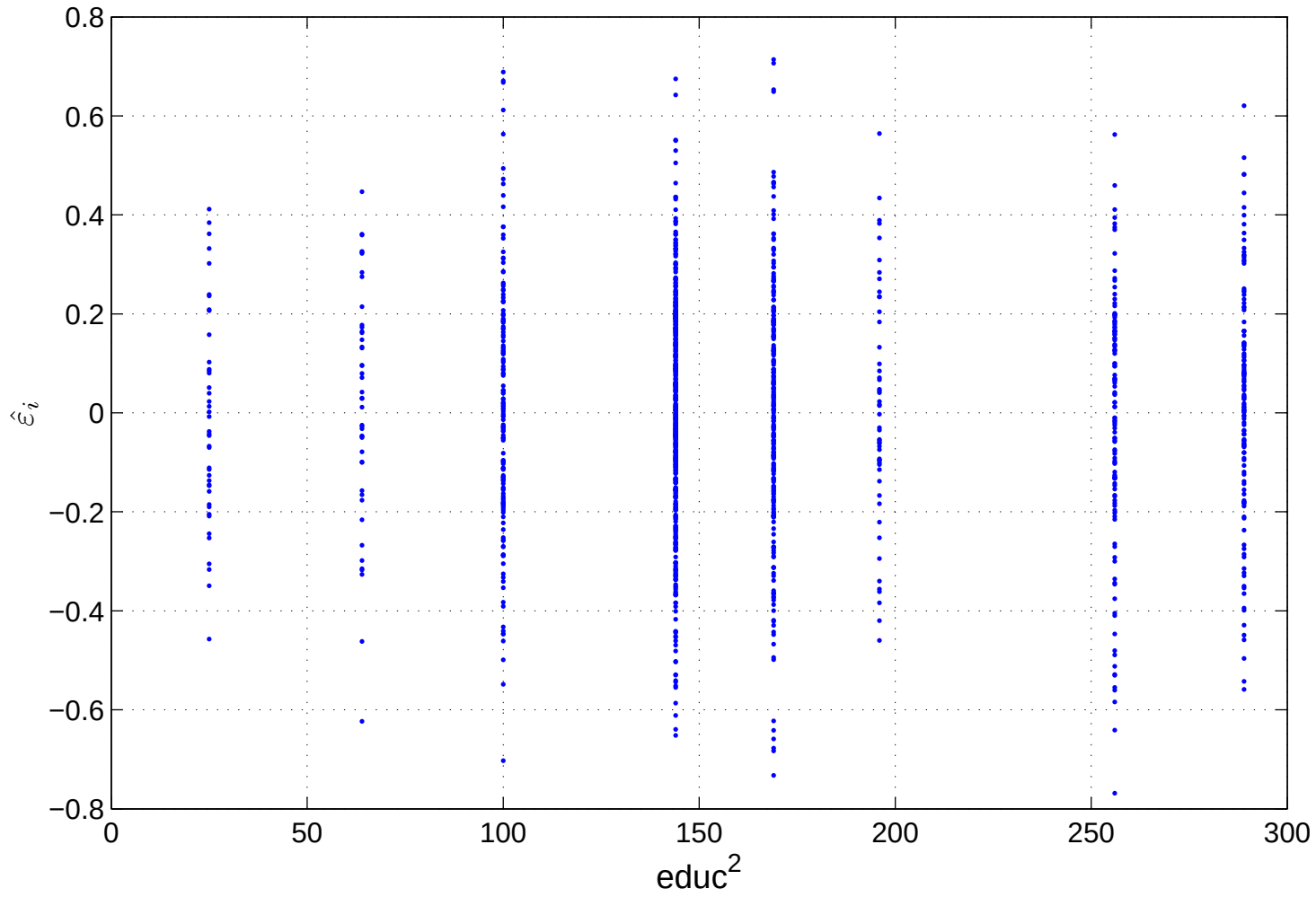
Остатки модели 3



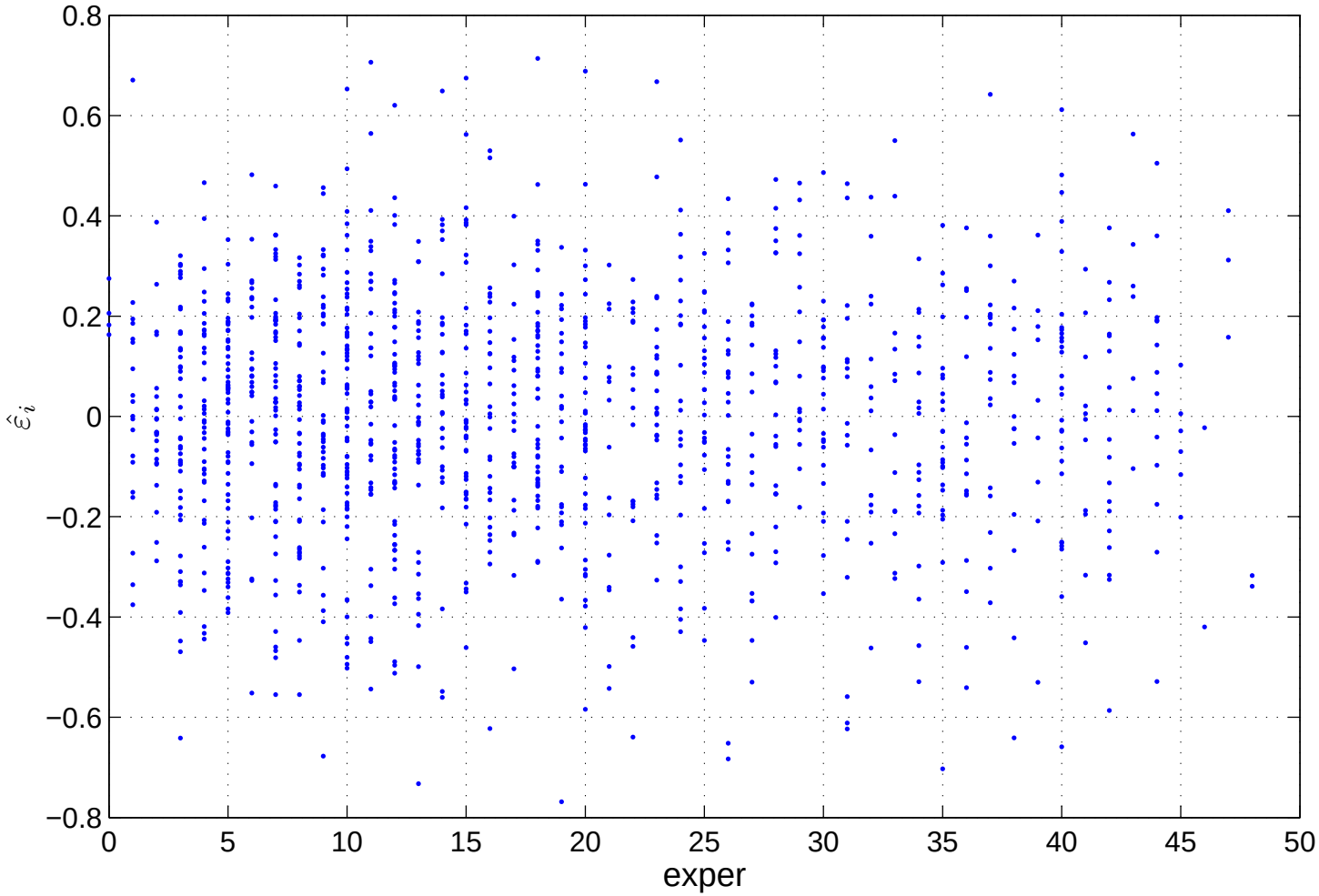
Остатки модели 3



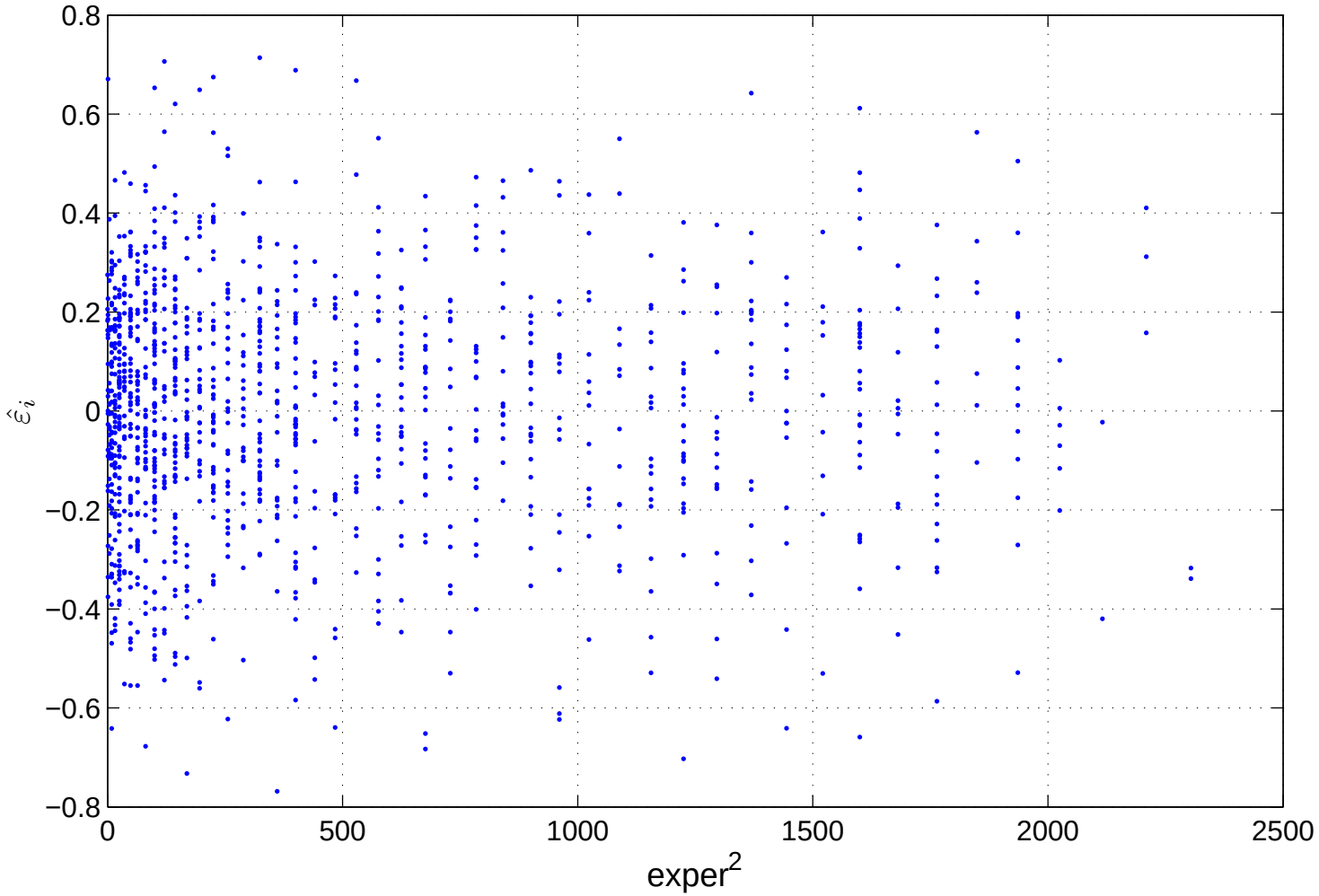
Остатки модели 3



Остатки модели 3

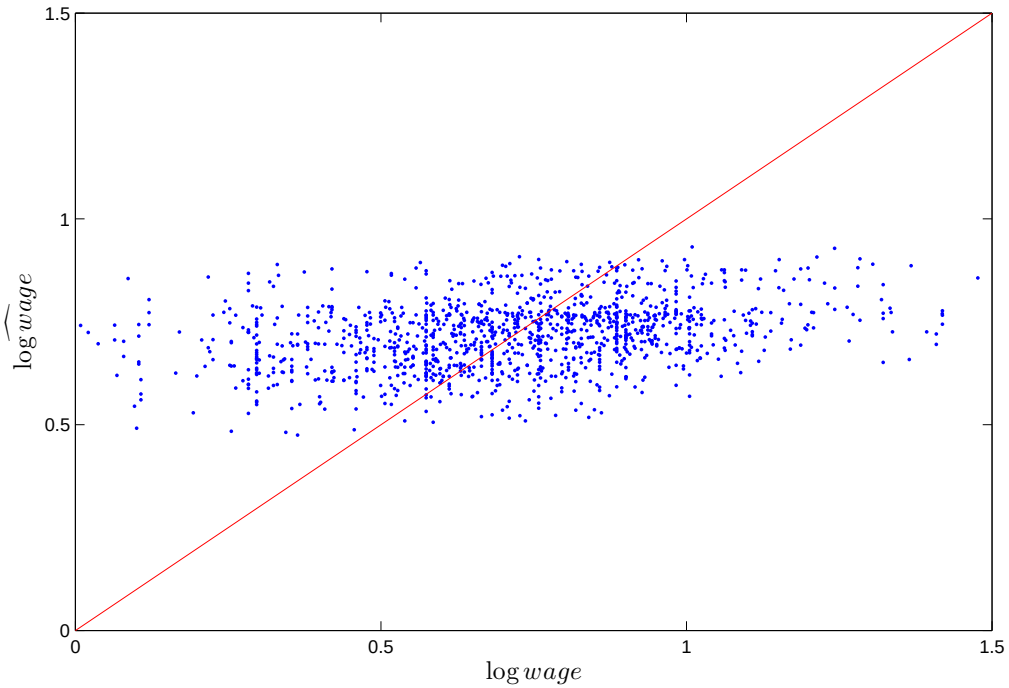


Остатки модели 3



Результат

Итоговая модель объясняет 10% вариации логарифма отклика:



С учётом дополнительных факторов, участники опроса с привлекательностью ниже среднего получают на 1% больше (95% доверительный интервал $(-3\%, 5.5\%)$), а с привлекательностью выше среднего — на 0.7% меньше (95% доверительный интервал $(-4.13\%, 2.58\%)$), причём в случае, если они состоят в профсоюзе, ещё на 4% меньше (95% доверительный интервал $(-11.25\%, 2.37\%)$).

Требования к решению задачи методом регрессии

Необходимо выполнить следующие операции:

- визуализация данных, анализ распределения признаков, оценка наличия выбросов;
- оценка необходимости преобразования отклика и его поиск методом Бокса-Кокса;
- отбор признаков (критерии Стьюдента, Фишера, пошаговая регрессия);
- визуальный анализ остатков;
- проверка гипотез об остатках: несмещённость, гомоскедастичность, некоррелированность, нормальность;
- анализ необходимости добавления взаимодействий и степеней признаков;
- расчёт расстояния Кука, возможное удаление выбросов, обновление модели;
- выводы.

Прикладная статистика
Регрессионный анализ, пример решения задачи.

Рябенко Евгений
riabenko.e@gmail.com