Постановка задачи

За каждый месяц 1991-2000 годов имеются следующие данные об электрозатратах и электропотреблении одного конкретного домохозяйства в Германии. Ежемесячно 1991-2000 проводились замеры затрат на электроэнергию в долларах, так же показаны следующие данные: средняя температура за месяц в фаренгейтах, CDD и HDD - погодные индексы (CDD - суммарное количество градусов, на которое средняя дневная температура выше 65°F; HDD - количество градусов, на которые средняя дневная температура ниже 65°F, взятое суммой за все дни месяца), количество живущих в доме человек, бинарные данные: указатель установки двух новых тепловых насосов, указатель установки нового счётчика, количество потребляемой электроэнергии выраженном в киловатт-часах.

Постановка задачи линейной регрессии

1…n — объекты;

$x_1 \ldots x_k$ — признаки;

y — зависимая переменная, потребление электроэнергии;

ε — ошибка;

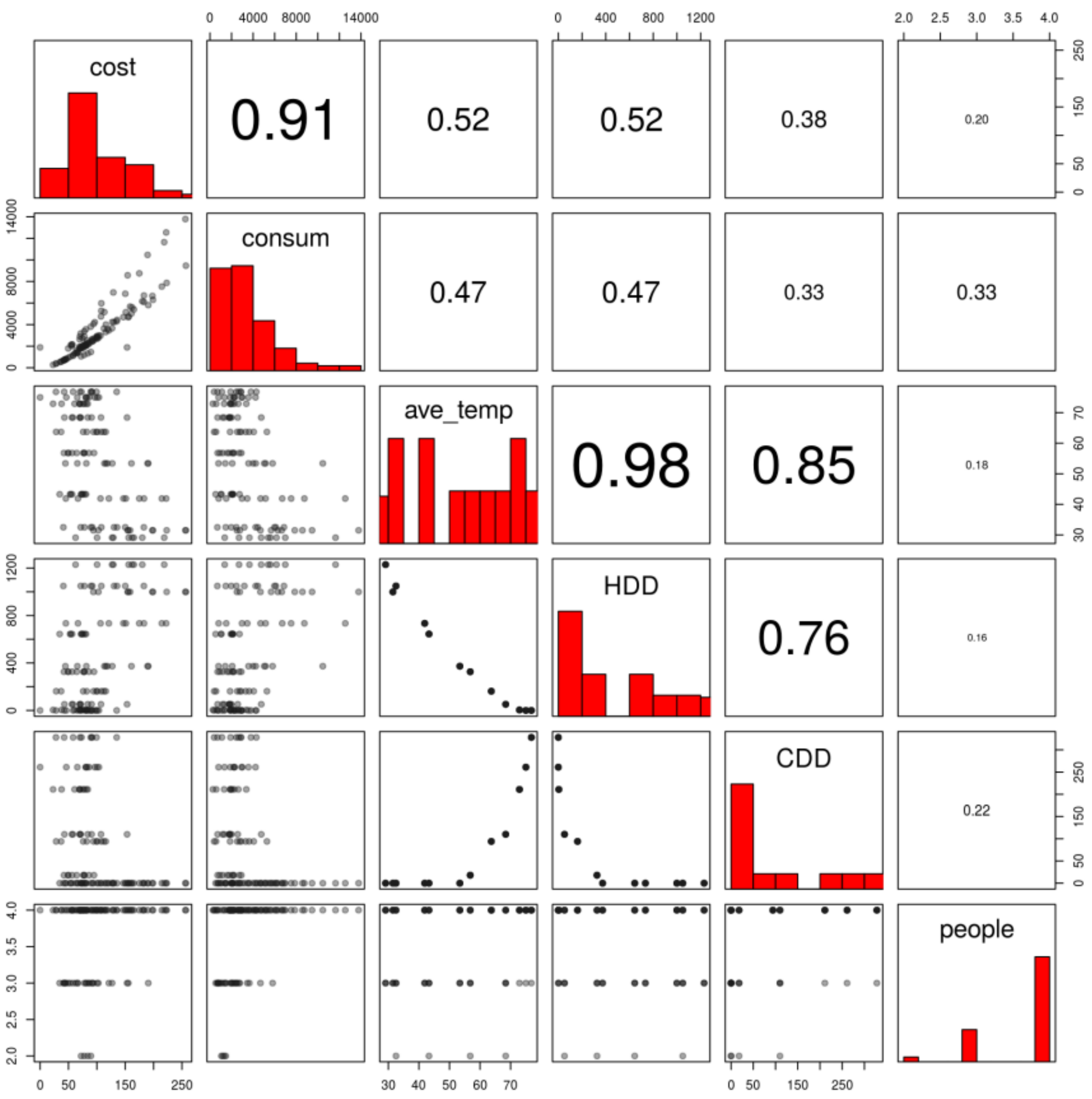
θ — неизвестные коэффициенты, которые хотим найти;

Хотим найти такую функцию f, что $y = f(x_1 \ldots x_k) + \varepsilon$;

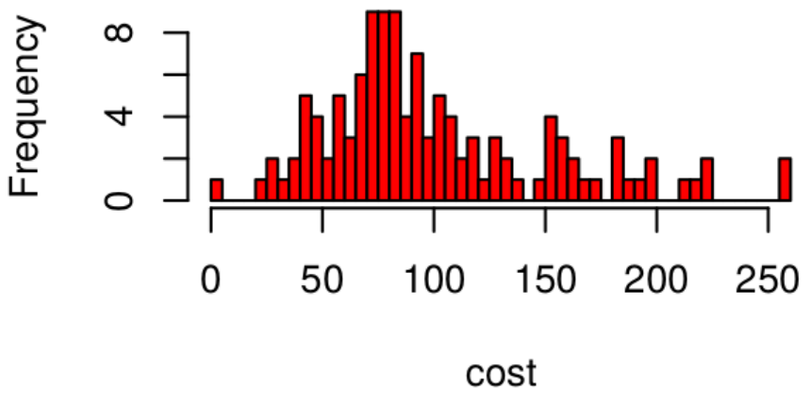
Множественная линейная регрессия:

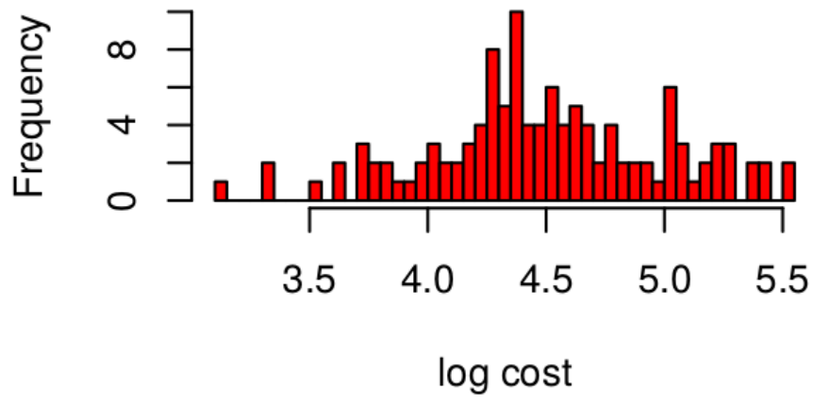
$$y_i = \theta_1 + \theta_2 x_{i,1} + \cdots + \theta_k x_{i,k-1} + \varepsilon_i$$

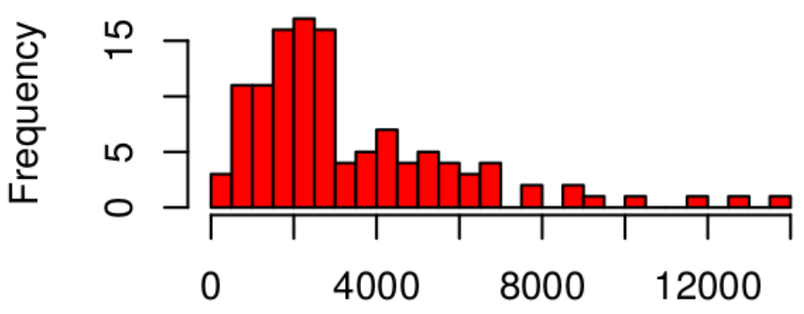
$\varepsilon = (\varepsilon_1 \ldots \varepsilon_k)^T$. $E\varepsilon = 0$. $K_\varepsilon = \sigma^2 I$.

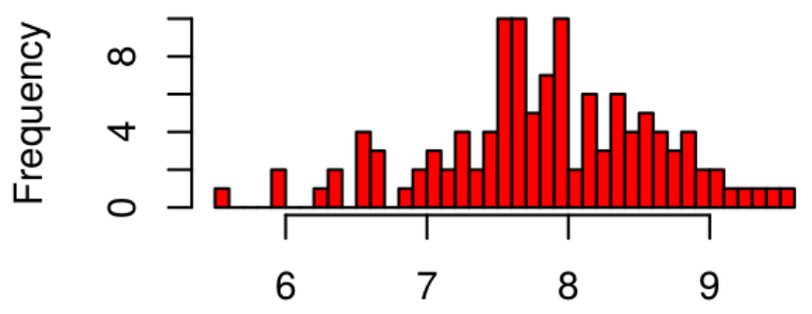

cost
consum
ave_temp
HDD
CDD
people
0.91
0.52
0.52
0.38
0.20
0.47
0.47
0.33
0.33
0.98
0.85
0.18
0.76
0.16
0.22

Решение









1. Исключим наблюдения, где затраты или потребление электричества равны 0.

2. $\frac{\max \text{cost}}{\min \text{cost}} = 11.2478109 > 10$ и $\frac{\max \text{consum}}{\min \text{consum}} = 52.2121212 > 10$, поэтому найдём преобразования откликов методом Бокса-Кокса:

Метод Бокса-Кокса

Пусть значения отклика $y_1, \ldots, y_n$ неотрицальны. Если $\max(y_i)/\min(y_i) > 10$, стоит посмотреть на возможность преобразования y. В каком виде его искать?

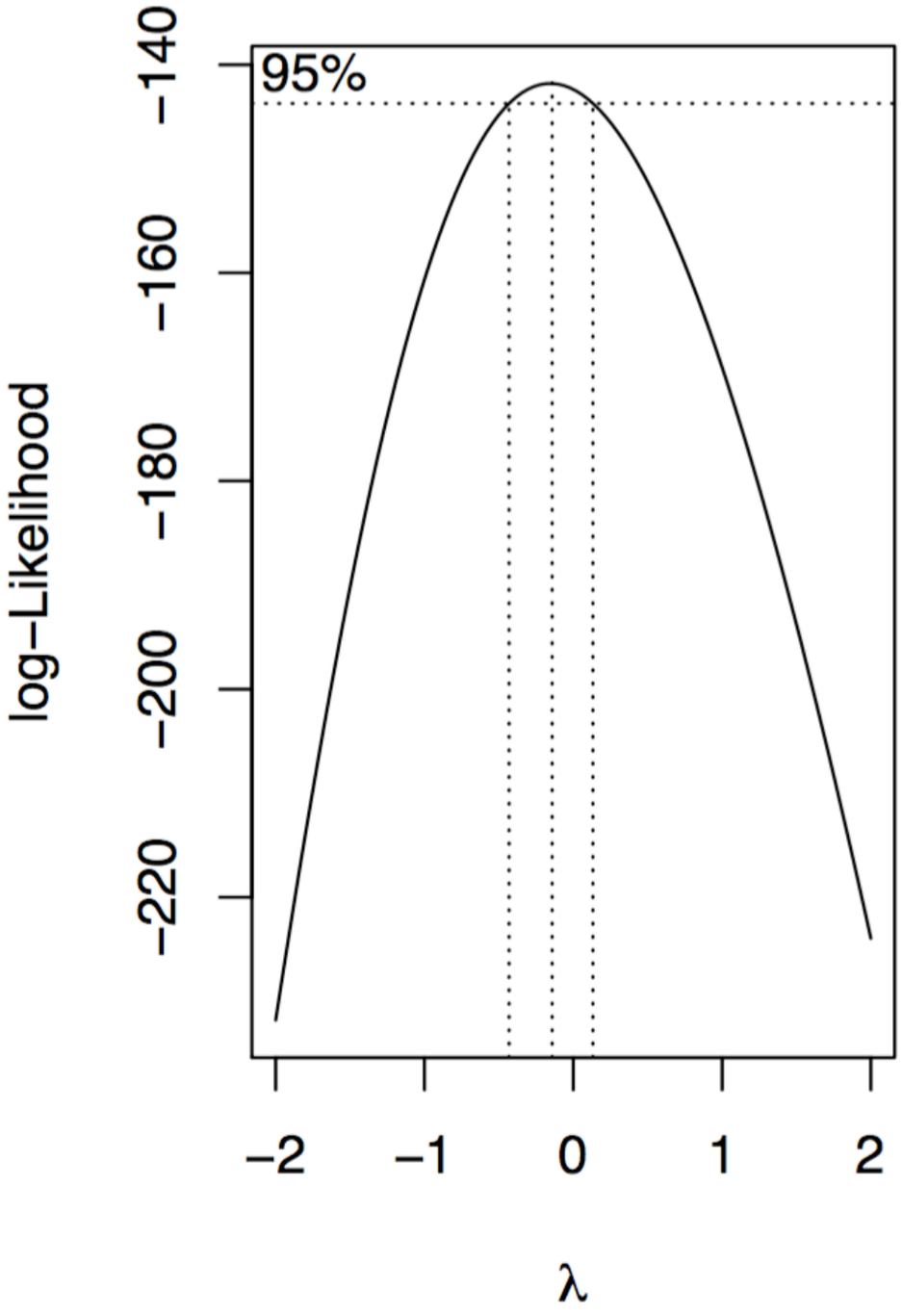
Рассмотривают преобразования вида y^λ, но они не имеет смысла при $\lambda = 0$. Вместо них можно рассмотреть группу преобразований:

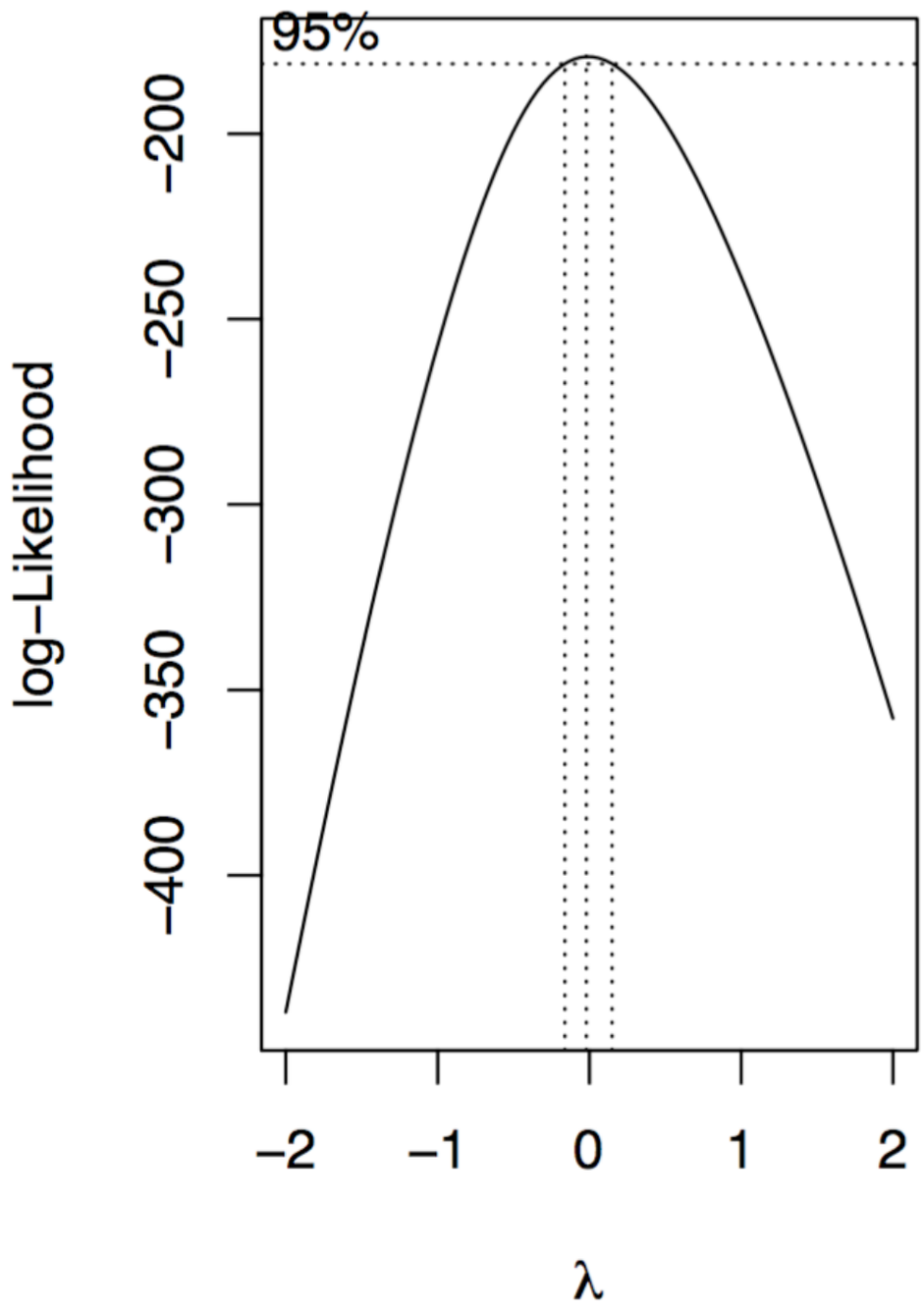
$$W = \begin{cases} (y^\lambda - 1)/\lambda & , \lambda \neq 0 \\ \ln y & , \lambda = 0 \end{cases}$$

Но они сильно варируется по λ. Вместо них рассмотрим семейство преобразований

$$V = \begin{cases} (y^\lambda - 1)/(\lambda \dot{y}^{\lambda-1}) & , \lambda \neq 0 \\ \dot{y} \ln y & , \lambda = 0 \end{cases}$$

Где $\dot{y} = (y_1, y_2, \ldots, y_n)^{1/n}$ — Среднее геометрическое наблюдений отклика.





В обоих случаях $\lambda = 0$ попадает в 95% доверительный интервал, поэтому будем строить регрессию логарифма отклика.

Модель потребления 1

Построим линейную модель для потребления по всем признакам, за исключением HDD.

Полученная модель линейной регрессии:

$$y = -0.030206 * ave\_temp + 0.001372 * CDD + 0.336694 * people + 1.459865 * counter - 0.767147 * pump\_1 - 0.782292 * pump\_2$$

Применив критерии Шапиро-Уилка, Бройша-Пагана и Стьюдента к остаткам модели, получаем p-value:

Критерий	p
Шапиро-Уилка	0.3584649
Стьюдента	1
Бройша-Пагана	0.0230271

Остатки нормальны и гетероскедастичны, поэтому для проверки несмещённости используем критерий Стьюдента

Условия для остатков:

$$E\varepsilon_i = 0,\ i = 1, \ldots, n$$
$$D\varepsilon_i = \sigma^2,\ i = 1, \ldots, n$$
$$\varepsilon_i \sim N(0, \sigma),\ i = 1, \ldots, n \text{ — нормально распределены}$$
$$\varepsilon_i,\ i = 1, \ldots, n \text{ — независимы}$$
$$\varepsilon_i = y_i - f_i,\ i = 1, \ldots, n$$

Для проверки гипотезы о нормальности остатков H_0: «остатки распределены нормально» — против альтернативы H_1: «остатки имеют не нормальное распределение» — используется критерий Шапиро-Уилка.

Критерий Бройша-Пагана применяется, если есть основания полагать, что дисперсия случайных ошибок может зависеть от некоторой совокупности переменных.

$H_0 : \sigma_1^2 = \cdot \cdot \cdot = \sigma_n^2 \Leftrightarrow$ остатки гомоскедастичны;

$H_1 : H_0$ — неправильна, т.е. остатки гетероскедантичны.

Метод включений-исключений

Удалим из модели 1 все признаки, кроме бинарных (модель 2).

Полученная модель линейной регрессии: y = 1.2234 ∗ counter − 0.9284 ∗ pump_1 − 0.7671 ∗ pump_2 Остатки модели 2:

Критерий	p
Шапиро-Уилка	0.0052393
Бройша-Пагана	0.4374139
Уилкоксона	0.6713462

Остатки не являются нормальными. Поэтому для проверки несмещённости используем критерий знаковых рангов Уилкоксона (метод часто используется для проверки гипотезы о равенстве средних двух независимых выборок) вместо критерия Стьюдента.
И гомоскедастичны, то есть дисперсии случайных ошибок модели постоянны.
Поэтому оценку значимости признаков будем делать с обычной оценкой дисперсии. Также будем делать поправку на множественность.

Будем добавлять новые регрессоры методом

включений. Т.к. признаков немного, переберем

также различные их взаимодействия.

Модель 3: + HDD · people

Полученная модель линейной регрессии:

$$y = 1.3087080 * counter - 0.8484593 * pump\_1 - 0.8133672 * pump\_2 + 0.0002379 * HDD * people$$

P-value теста Бройша-Пагана для модели 3 составляет 0.707, т.е. остатки гетероскедастичны. Сравним модели 2 и 3 с помощью критерия Вальда. Критерий Вальда проверяет гипотезу H_0 о равенстве нулю коэффициентов при параметрах, которые появились в новой модели, против противоположной альтернативы H_1. Если p-value по критерию Вальда меньше выбранного уровня значимости, отвергаем H_0, т.е. считаем новую модель лучше предыдущей.

$p = 2.2 \cdot 10^{-16} < 0.05$ — модель 3 лучше модели 2.

Модель 4: +people * counter

Полученная модель линейной регрессии:

$y = -1.3087080 * counter - 0.8487730 * pump\_1 - 0.8131858 * pump\_2$

$-0.0002369 * HDD * people + 0.6136127 * people * counter$

Сравним ее с моделью 3 при помощи критерия Вальда.

$p = 2{,}7 * 10^{-8}$ По тем же соображениям видим, что модель 4 лучше, чем модель 3.

По тем же соображениям видим, что модель 4 лучше, чем модель 3.

Модель 5: +CDD * (1 − pump_1)

Полученная модель линейной регрессии: $y = -0.3906500 * counter - 0.7336461 * pump\_1 - 0.8212860 *$

$pump\_2 + 0.0002786 * HDD * people + 0.6094133 * people * counter + 0.0011709 * CDD * (1 - pump_1)$ Сравним ее с

предыдущей моделью:

p = 0.02251, Модель 5 лучше, чем модель 4. И это пока нас устраивает.

Модель 6: ave_temp : (1 − counter)

Полученная модель линейной регрессии: y = −1.21775350 ∗ counter − 0.7165238 ∗ pump_1 −0.8714436∗pump_2 +0.0001565∗HDD∗people+0.6217141∗people∗counter+0.0016678∗CDD∗

$(1 - pump_1) - 0.0149175 * ave\_temp * (1 - counter)$ Применим тест Вальда:

p = 0.005168, Опять же, модель стала лучше.

Модель 7: people : (1 − pump_2)

Полученная модель линейной регрессии: y = −1.5738140 ∗ counter − 0.5253195 ∗ pump_1 +0.5579918∗pump_2 +0.0001105∗HDD∗people+0.6263581∗people∗counter+0.0019657∗CDD∗ $(1 - pump_1)$ −

0.0214221 ∗ ave_temp ∗ (1 − counter) + 0.4067542 ∗ people ∗ (1 − pump_2)

Сравним с прошлой моделью: p = 0.00837, Старая модель хуже. После 7-й модели провела все возможные Вальд-тесты для новых регрессоров. Минимальное p-value при добавлении еще

одного признака составляет 0.334 — больше выбранного уровня значимости, следовательно, нет смысла в дополнительных регрессорах.

Остатки модели 7:

Критерий	p
Шапиро-Уилка	0.4897525
Бройша-Пагана	0.0931999

Остатки модели 7 нормальны и гетероскедастичны.

Удалим признак pump_2:

Новая модель лучше, поэтому перейдем к этой модели 8.

Ее остатки:

Критерий	p
Шапиро-Уилка	0.380842
Бройша-Пагана	0.0936944

Остатки модели 8 нормальны и гетероскедастичны.

Получившиеся уравнение линейной регрессии :

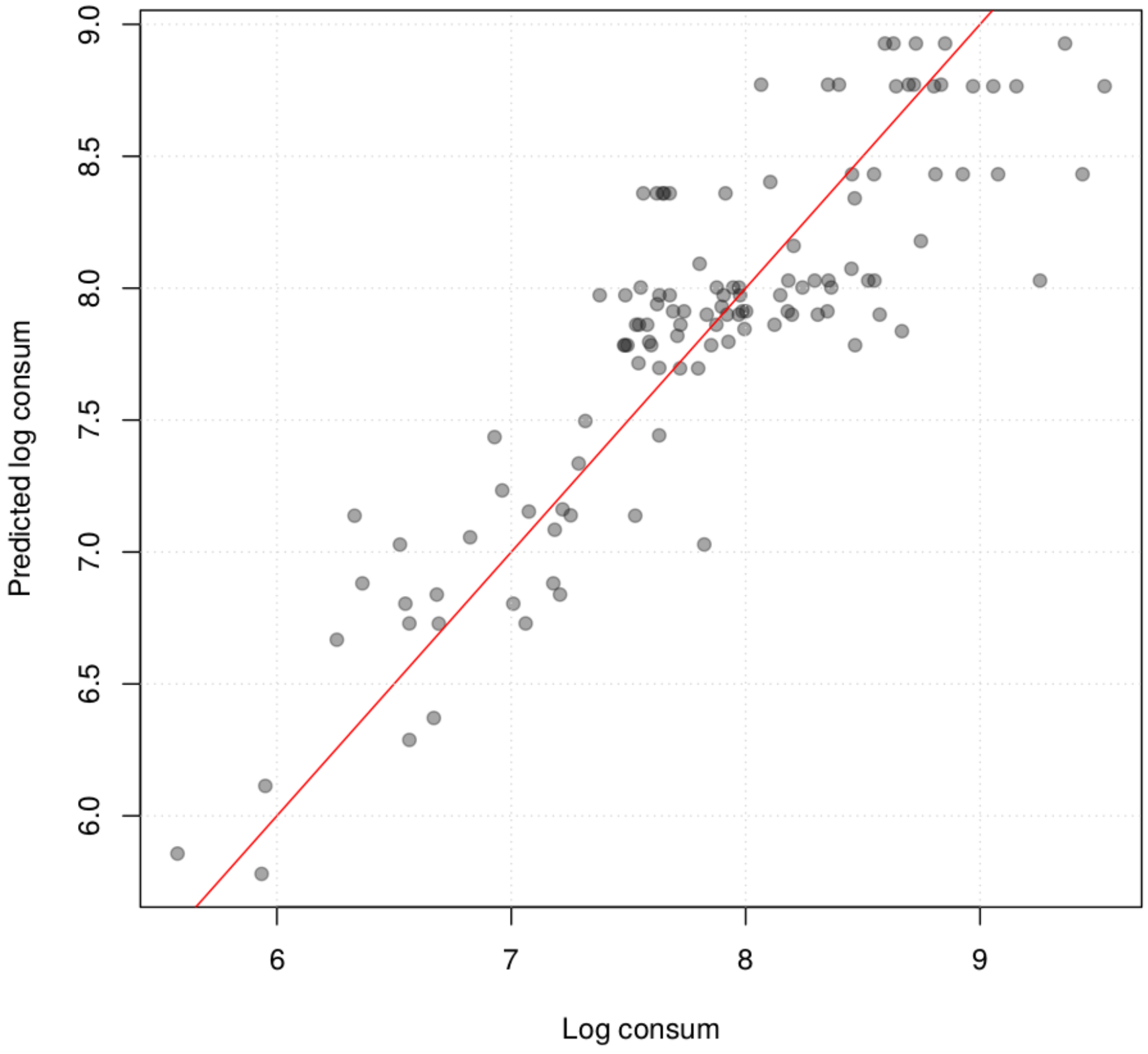
$$y = -1.435 * counter - 5.867 * 10^{-1} * pump_1 + 1.26 * 10^{-4} * HDD * people +$$

$$6.248 * 10^{-1} people * counter + 2.56 * 10^{-1} * people * (1 - pump_2) -$$

$$1.92 * 10^{-2} * ave\_temp * (1 - counter) + 1.865 * 10^{-3} * CDD * (1 - pump_1)$$

Интерпретация коэффициентов

Итоговая модель для потребления электричества объясняет 76% вариации логарифма отклика:



При интересующих нас факторах стоят следующие коэффициенты:

counter	pump_1	people:I(counter)
-1.435213114	-0.586745553	0.624797044
people:I(1 - pump_2)	ave_temp:I(1 - counter)	CDD:I(1 - pump_1)
0.256041692	-0.019230974	0.001864978

95% доверительные интервалы:

	2.5 %	97.5 %
counter	-2.8065716885	-0.063854540
pump_1	-0.8300216211	-0.343469486
people:I(counter)	0.2493682557	1.000225833
people:I(1 - pump_2)	0.1706818997	0.341401484
ave_temp:I(1 - counter)	-0.0342533947	-0.004208554
CDD:I(1 - pump_1)	0.0006819577	0.003047998

Вывод :

Таким образом, с учётом того, что изначально мы перешли к логарифму, то коэффициенты будут браться как степень экспоненты и дальше будем их интерпретировать.

- установка нового счетчика уменьшила потребление в $e^{1.43} = 4.2$ раза,
- установка первого теплового насоса уменьшила расходы на 44%,
- установка второго теплового насоса сама по себе не повлияла на потребление,
- с установкой нового счетчика каждый дополнительный человек в доме увеличивал потребление на 87%,
- пока не был установлен второй насос, каждый человек увеличивал потребление на 29%,
- до установки нового счетчика, каждый градус средней температуры уменьшал потребление на 1.9%
- до установки первого насоса, каждый градус CDD повышал потребление на 0.19%.