Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

КУРСОВАЯ РАБОТА

на тему

«Построение математических моделей экономических процессов методами регрессионного анализа»

Выполнила студентка группы БПМИ145, 2 курса,
Чеснокова Полина Владимировна

Научный руководитель:
к.ф.-м.н., доцент
Горяинова Елена Рудольфовна

Москва 2016

1. Введение

В современном мире все чаще используется анализ и обработка статистической информации. В данной курсовой работе будет рассмотрена проблема построения линейной регресии и ее улучшение на основании данных, связанных с электрозатратами.
Будет рассмотрена практическая задача: как скажется установка нового оборудования на потребление электричества, и построена оптимальная модель линейной регрессии, прогнозирующая электропотребление в зависимости от других входных данных.

2. План

- визуализация данных, анализ распределения признаков, оценка наличия выбросов;
- оценка необходимости преобразования отклика и его поиск методом Бокса-Кокса;
- визуальный анализ остатков;
- проверка нормальности, несмещённости и гомоскедатичности остатков;
- множественная проверка гипотез и отбор признаков с учётом гетероскедастичности;
- анализ необходимости добавления взаимодействий и квадратов признаков;
- вычисление расстояний Кука, возможное удаление выбросов, улучшение модели;
- выводы.

3.Постановка задачи

За каждый месяц 1991-2000 годов имеются следующие данные, которые находятся по ссылке (https://www.dropbox.com/s/za6mf4y9wzobukp/Electricity.csv?dl=0) об электрозатратах и электропотребление одного конкретного домохозяйства в Германии. Каждый месяц 1991-2000 проводильсь замеры затрат на электроэнергию в долларах, так же приведены сделующие данные: средняя температура за месяц в фарингейтах, погодные индексы CDD и HDD (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,индикаторы установки двух новых тепловых насосов, индикатор установки нового счётчика, объём потребления электроэнергии выраженном в киловатт-часах.

- затраты на электроэнергию, $;
- среднемесячная температура по данным последних 30 лет;
- погодный индекс CDD (Cooling Degree Day), усредненный за 30 лет по месяцам;
- погодный индекс HDD (Heating Degree Day), усредненный за 30 лет по месяцам;
- число проживающих в доме членов семьи;
- бинарные признаки: индикатор установки нового счётчика, индикаторы установки двух новых тепловых насосов;
- объём потребления электроэнергии, кВт · ч.

Требуется построить оптимальную модель линейной регрессии.

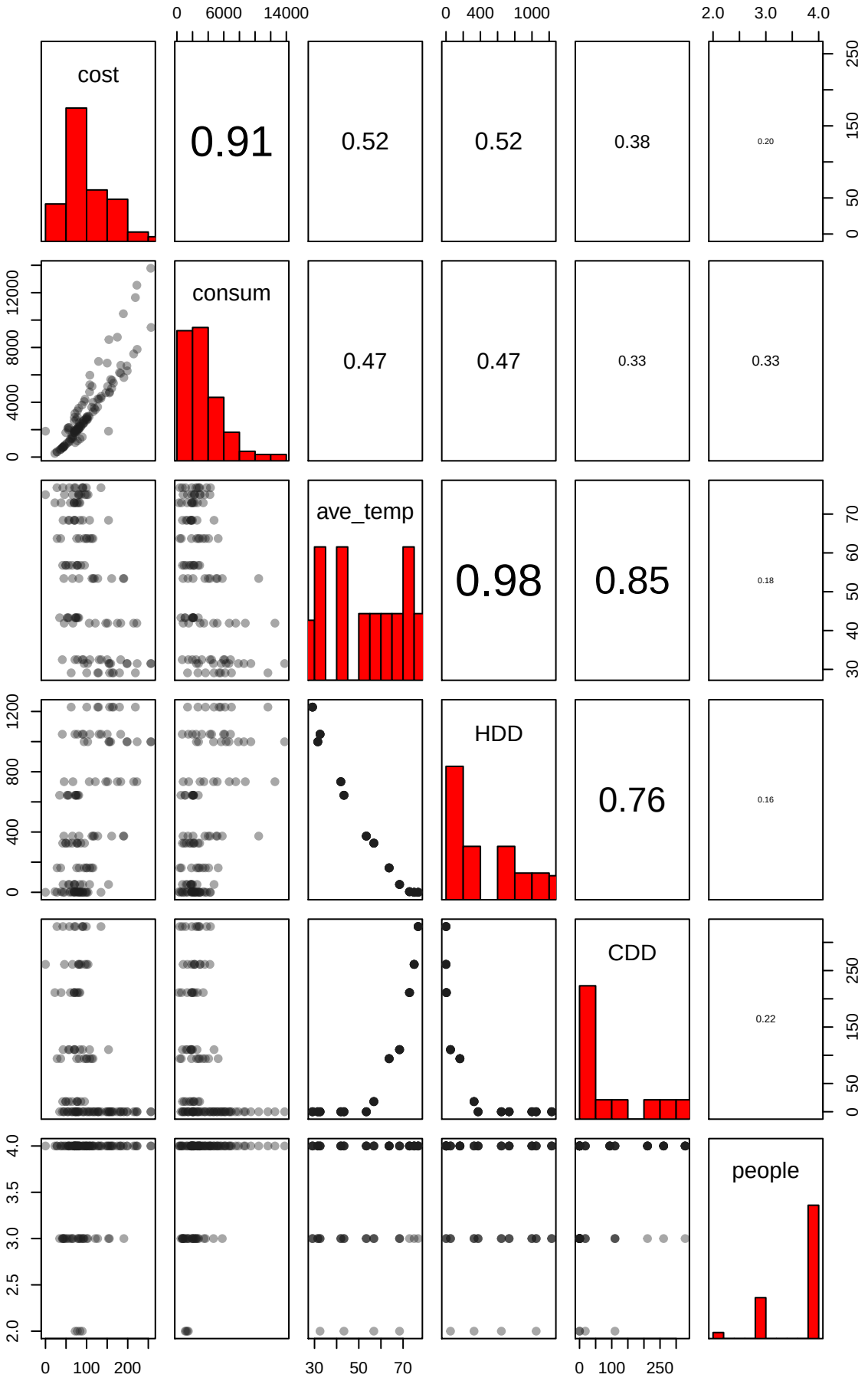
Постановка задачи линейной регрессии:

1, . . . , n — объекты;
x1, . . . , xk, y — признаки, значения которых измеряются на объектах;
x1, . . . , xk — объясняющие переменные (предикторы, регрессоры,факторы, признаки);

y — зависимая переменная, отклик.
Хотим найти такую функцию f, что $y \approx f(x1, ..., xk) + \epsilon$; $\arg\min_f E(y - f(x_1, \ldots, x_k))^2 = E(y|x_1, \ldots, x_k)$.
E (y |x1, . . . , xk) = f (x1, . . . , xk) — модель регрессии;
E (y |x1, . . . , xk) = $(\beta_0 + \sum_{j=1}^{k} \beta_j x_j)$ — модель линейной регрессии.

В задаче регрессии требуется восстановить значение целевого вектора на основе заданных признаков и при ее решении имеет значение качество входных данных. Модель называется устойчивой, если малые изменения вектора параметров приводят к малым изменениям целевого вектора. Мультиколлинеарность — частичная линейная зависимость между признаками — приводит к значительному снижению устойчивости модели.[1] Для решения проблемы мультиколлинеарности предлагается применить отбор признаков.

Попарные диаграммы рассеяния всех количественных признаков (на графиге по осям: расходы, постебление, средняя температура, HDD, CDD, количество человек в семье. Сверху вниз и слева направо соответсвенно):



Как видим, расходы и потребление электроэнергии не являются линейно зависимыми (потому что они зависят от цены на электричество, которая нам неизвестна), но приблизительно лежат на двух прямых. Также, средняя температура, HDD и CDD принимают по 12 значений (на каждый месяц), причем температура и HDD имеют отрицательную корреляцию, а температура и CDD — положительную.

Постановка задачи отбора признаков

Предлагается подход шаговой регрессии

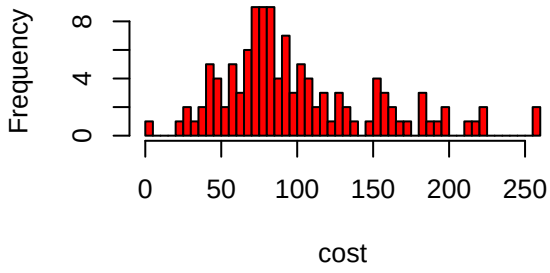
В статистике, пошаговая регрессия включает в себя модели регрессии, в которых выбор прогнозирующих переменных осуществляется автоматической процедурой. Исследуется зависимость информативности отобранных признаков отпараметров шаговой регрессии. В вычислительном эксперименте алгоритм тестируется на данных потребителей электроэнергии, а так же о данных о погодных условиях.Как правило, это принимает форму последовательности F-тестов или t - тестов. Алгоритм:
* Шаг 0. Настраивается модель с единственной константой, а также все модели с одной переменной. Рассчитывается F-статистика каждой модели и достигаемый начальный уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная X_{e1} включается в модель, если этот достигаемый уровень значимости меньше порогового значения $p_E = 0.05$.
* Шаг 1. Рассчитывается F-статистика и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых X_{e1}. Аналогично принимается решение о включении X_{e2}.
* Шаг 2. Если добавленная переменная X_{e2}, может оказаться, что, X_{e1} уже не нужна. В общем случае просчитываются все возможные варианты исключения одной переменной, рассматривается вариант с наибольшим достигаемым уровнем значимости, соответствующая переменная исключается, если он превосходит пороговое значение $p_R = 0.1$.
...

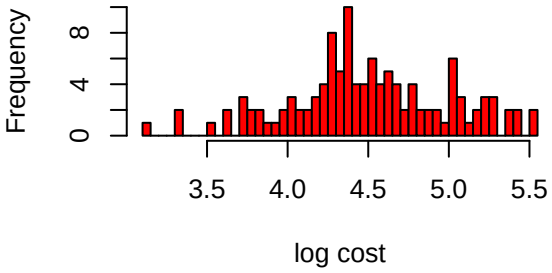
Способ проверки на наличие ошибок в моделях, созданных методом шаговой регресии, нужно использовать абсолютно новые данные, а не то, с помощью которых мы строили данную модель. Для этого для построения модели используется, например 70% а остальные 30% используется для проверки точности модели. Точность затем часто измеряется как фактическая стандартная ошибка, или средняя ошибка между предсказанным значением и фактическим значением. Этот метод является особенно ценным, когда данные собираются в различных условиях или когда модели предполагаются обобщению.

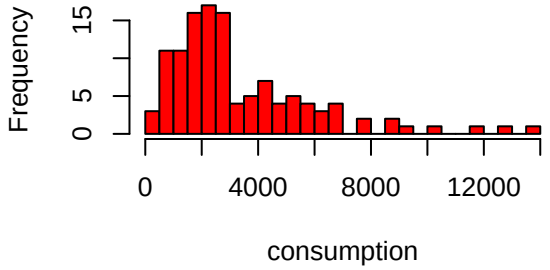
4.Решение

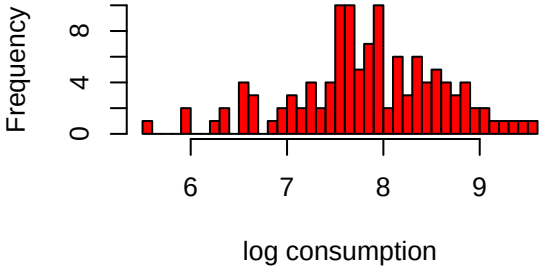
Предобработка

Распределение значений откликов:









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 $y_1, ..., y_n$ 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 оно не имеет смысла при $\lambda = 0$. Вместо него можно рассмотреть группу преобразований:

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

Но оно сильно варируется по λ. Вместо него рассмотрим семейство преобразований

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

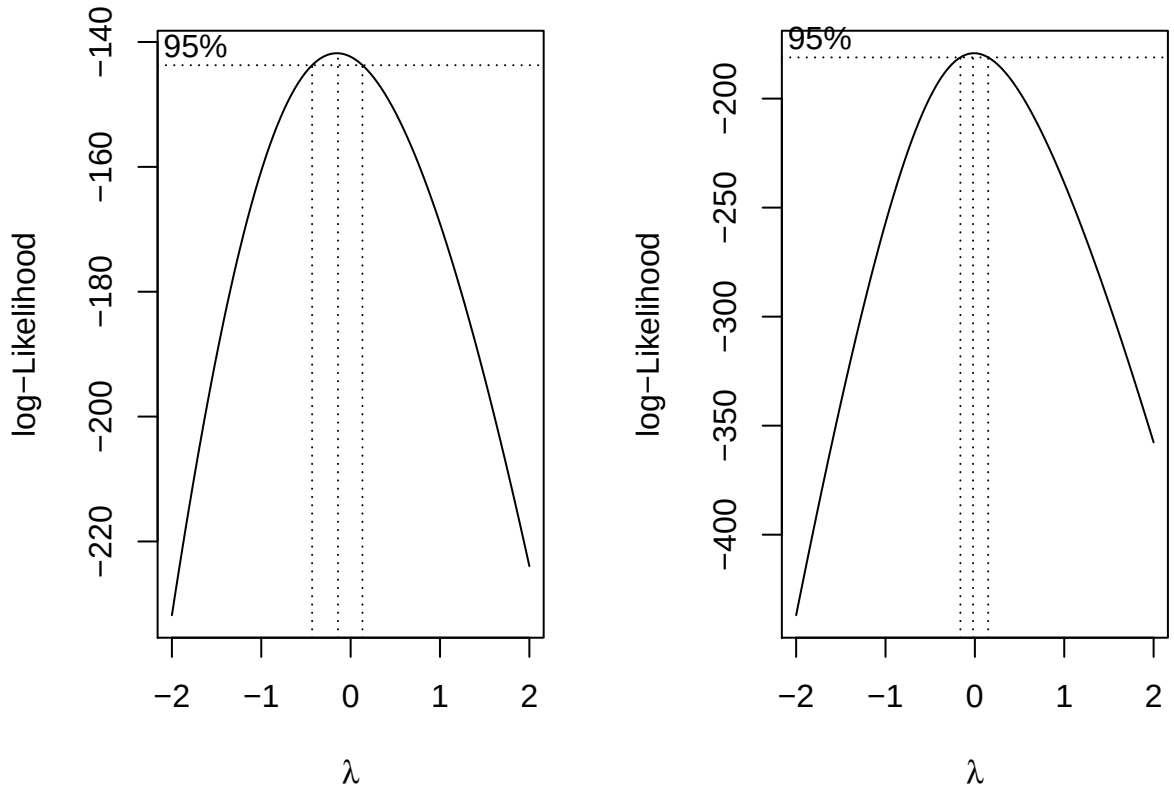
Где

$$\dot{y} = (y_1 y_2 \dots y_n)^{1/n}$$

- Среднее геометрическое наблюдение отклика

Алгоритм выбора λ:

- выбирается набор λ в неком интервале
- выполняется преобразование $V(\lambda)$ для каждого значения λ, строится регрессия V на X, вычисляется RSS(λ) — остаточная сумма квадратов;
- строится график зависимости RSS (λ) от λ, по нему выбирается оптимальное значение λ;
- выбирается ближайшее к оптимальному удобное значение λ;
- строится окончательная регрессионная модель с откликом y^λ или $\ln y$.



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

Модель потребления 1

Построим линейную модель для затрат по всем признакам.

Её остатки:

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

Для того, чтобы получить информацию о пригодности той модели многомерной линейной регрессии, которую мы постороили исследуют регрессионные остатки. При условии, что регрессионная модель, которую мы выбрали адекватно описывает истинную зависимость, то должны быть выполнены следующие условия для остатков:
$E\varepsilon_i = 0,\ i = 1, \cdots, n$
$D\varepsilon_i = \sigma^2,\ i = 1, \cdots, n$
$\varepsilon_i \sim N(0, \sigma),\ i = 1, \cdots, n$ - нормально распределены
$\varepsilon_i,\ i = 1, \cdots, n$ - независимы
$\varepsilon_i = y_i - f_i,\ i = 1, \cdots, n$

Чтобы проверить выполнение этих условий используются разные критерии, например, для проверки гипотезы о нормальности остатков: H_0: «случайная величина X распределена нормально» используется критерий Критерий Шапиро-Уилка.

Для проверки несмещённости используем критерий Стьюдента: H_0: E(X)=m о равенстве математического ожидания E(X) некоторому известному значению m.

Критерий Бройша-Пагана используем для проверки наличия гетероскедастичности : $H_0: \ \gamma_2 = \cdots = \gamma_p = 0 \Leftrightarrow \sigma_1^2 = \cdots = \sigma_n^2 \Leftrightarrow$ остатки гомоскедастичны;

Метод включений-исключений

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

Критерий	p
Шапиро-Уилка	0.0052393
Бройша-Пагана	0.4374139

Ошибки не являются нормальными. Возникает проблема : если остатки не являются нормально распределенными, то в качестве зависимой переменной, или, по крайней мере одна объясняющая переменная может иметь неправильную функциональную форму, или важные переменные могут отсутствовать.
Поэтому для проверки несмещённости используем критерий знаковых рангов Уилкоксона.
И гомоскедастичны, однородность наблюдений, то есть постоянство дисперсии случайных ошибок модели.
Поэтому оценку значимости признаков будем делать с обычной оценкой дисперсии. Также будем делать поправку на множественность.
Алгоритм:
1. Составить модель испытуемых в произвольном порядке.
2. Посчитать разность между индивидуальными значениями во втором и первом измерениях. Определить, что будет считаться нормальным сдвигом.
3. Согласно алгоритму ранжирования, проранжировать абсолютные величины разностей, начисляя меньшему значению меньший ранг, и проверить совпадение полученной суммы рангов с расчетной.
4. Отметить каким-либо способом ранги, соответствующие сдвигам в нетипичном направлении. Подсчитать их сумму T.
5. Определить критические значения T для данного объема выборки. Если T-эмп. меньше или равен T-кр. – сдвиг в «типичную» сторону достоверно преобладает.
По сути оцениваются знаки значений, полученных вычитанием ряда значений одного измерения из другого. Если в результате количество уменьшившихся значений примерно равно количеству увеличившихся, то подтверждаем гипотезу о нулевой медиане.

```
##
## Call:
## lm(formula = log_consum ~ counter + pump_1 + pump_2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9566 -0.4909 -0.0353  0.3734  1.4594
##
## Coefficients:
##             Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)  8.217e+00   6.834e-02   1.202e+02  6.355e-122
## counter      1.223e+00   2.321e-01   5.270e+00   6.532e-07
## pump_1       -9.284e-01   1.430e-01  -6.490e+00   2.308e-09
## pump_2       -7.671e-01   2.213e-01  -3.467e+00   7.426e-04
##             Adjusted p-value
## (Intercept)               NA
```

```
## counter                0.000
## pump_1                 0.000
## pump_2                 0.002
##
## Residual standard error: 0.5759 on 114 degrees of freedom
## Multiple R-squared:  0.4735, Adjusted R-squared:  0.4597
## F-statistic: 34.18 on 3 and 114 DF,  p-value: 7.813e-16
```

Будем добавлять новые регрессоры методом включений. Т.к. признаков немного, переберем также различные их взаимодействия.

Модель 3: + HDD:people, p-value = 6.6×10^{-16}, AIC = -192.7. где p-value, такое, что если p-value меньше альфы, которая выбирается (уровень значимости), то гипотеза о нормальности распределения отвергается.

P-value теста Бройша-Пагана для модели 3 составляет 0.707, т.е. остатки гетероскедастичны. Сравним модели 2 и 3 с помощью критерия Вальда. То есть будем оценивать модели исходя из их наихудшести. То есть, лучшее (оптимальное) решение является тем , чей худший результат , по крайней мере так же хорош , как худший результат каких - либо других решений

```
## Wald test
##
## Model 1: log_consum ~ HDD:people + counter + pump_1 + pump_2
## Model 2: log_consum ~ counter + pump_1 + pump_2
##   Res.Df Df      F    Pr(>F)
## 1    113
## 2    114 -1 120.26 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Новая модель значимо лучше.

Модель 4: + people:counter, p-value = 0.002, AIC = -200.4. Сравним ее с моделью 3 при помощи критерия Вальда.

```
## Wald test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + counter + pump_1 +
##     pump_2
## Model 2: log_consum ~ HDD:people + counter + pump_1 + pump_2
##   Res.Df Df      F    Pr(>F)
## 1    112
## 2    113 -1 35.798 2.679e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Как видим, модель 4 лучше, чем модель 3.

Модель 5: + CDD:(1 - pump_1), p-value = 0.037, AIC = -203.1. Сравним ее с предыдущей моделью:

```
## Wald test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + CDD:I(1 - pump_1) +
##     counter + pump_1 + pump_2
## Model 2: log_consum ~ HDD:people + people:I(counter) + counter + pump_1 +
##     pump_2
##   Res.Df Df      F  Pr(>F)
## 1    111
## 2    112 -1 5.3545 0.02251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Модель 5 немного лучше, чем модель 4.

Модель 6: ave_temp:(1 - counter), p-value = 0.051, AIC = -205.2. Тест Вальда:

```
## Wald test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + ave_temp:I(1 -
##     counter) + CDD:I(1 - pump_1) + counter + pump_1 + pump_2
## Model 2: log_consum ~ HDD:people + people:I(counter) + CDD:I(1 - pump_1) +
##     counter + pump_1 + pump_2
##   Res.Df Df      F    Pr(>F)
## 1    110
## 2    111 -1 8.1425 0.005168 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Опять же, модель стала лучше.

Модель 7: people:(1 - pump_2), p-value = 0.036, AIC = -207.9. Сравним с прошлой моделью:

```
## Wald test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + people:I(1 - pump_2) +
##     ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter + pump_1 +
##     pump_2
## Model 2: log_consum ~ HDD:people + people:I(counter) + ave_temp:I(1 -
##     counter) + CDD:I(1 - pump_1) + counter + pump_1 + pump_2
##   Res.Df Df      F  Pr(>F)
## 1    109
## 2    110 -1 7.2129 0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Старая модель хуже.

Минимальное p-value при добавлении еще одного признака составляет 0.334, AIC увеличился (-207). Следовательно, нет смысла в дополнительных регрессорах.

Остатки модели 7:

Критерий	p
Шапиро-Уилка	0.4897525
Стьюдента	1
Бройша-Пагана	0.0931999

нормальны, поэтому для проверки несмещённости используем критерий Стьюдента

```
##
## Call:
## lm(formula = log_consum ~ HDD:people + people:I(counter) + people:I(1 -
##     pump_2) + ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter +
##     pump_1 + pump_2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85593 -0.27960 -0.01509  0.26145  1.22897
##
## Coefficients:
##                              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)                 7.379e+00  7.626e-01  9.675e+00  2.353e-16    NA
## counter                    -1.574e+00  7.140e-01 -2.204e+00  2.961e-02 0.008
## pump_1                     -5.253e-01  1.445e-01 -3.634e+00  4.268e-04 0.002
## pump_2                      5.580e-01  6.904e-01  8.083e-01  4.207e-01 0.824
## HDD:people                  1.105e-04  7.145e-05  1.546e+00  1.250e-01 0.104
## people:I(counter)           6.264e-01  1.898e-01  3.301e+00  1.303e-03 0.000
## people:I(1 - pump_2)        4.068e-01  1.914e-01  2.125e+00  3.582e-02 0.048
## ave_temp:I(1 - counter) -2.142e-02  8.062e-03 -2.657e+00  9.062e-03 0.001
## CDD:I(1 - pump_1)           1.966e-03  6.107e-04  3.219e+00  1.697e-03 0.002
##
## Residual standard error: 0.3994 on 109 degrees of freedom
## Multiple R-squared:  0.7578, Adjusted R-squared:   0.74
## F-statistic: 42.64 on 8 and 109 DF,  p-value: < 2.2e-16
```

Удалим признак pump_2:

```
## Wald test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + people:I(1 - pump_2) +
##     ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter + pump_1 +
##     pump_2
## Model 2: log_consum ~ HDD:people + people:I(counter) + people:I(1 - pump_2) +
##     ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter + pump_1
##   Res.Df Df      F Pr(>F)
## 1    109
## 2    110 -1 1.0847 0.2999
```

Новая модель лучше, поэтому перейдем к этой модели (№8).

Ее остатки:

Критерий	p
Шапиро-Уилка	0.380842
Стьюдента	1
Бройша-Пагана	0.0936944

нормальны, поэтому для проверки несмещённости используем критерий Стьюдента

```
##
## Call:
## lm(formula = log_consum ~ HDD:people + people:I(counter) + people:I(1 -
##     pump_2) + ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter +
##     pump_1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80613 -0.28151 -0.03142  0.26475  1.22684
##
## Coefficients:
##                              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)                 7.844e+00  5.000e-01  1.569e+01  7.973e-30    NA
## counter                    -1.435e+00  6.920e-01 -2.074e+00  4.041e-02 0.008
## pump_1                      -5.867e-01  1.228e-01 -4.780e+00  5.470e-06 0.000
## HDD:people                   1.260e-04  6.872e-05  1.833e+00  6.955e-02 0.027
## people:I(counter)            6.248e-01  1.894e-01  3.298e+00  1.312e-03 0.000
## people:I(1 - pump_2)         2.560e-01  4.307e-02  5.944e+00  3.327e-08 0.000
## ave_temp:I(1 - counter) -1.923e-02  7.580e-03 -2.537e+00  1.258e-02 0.001
## CDD:I(1 - pump_1)            1.865e-03  5.970e-04  3.124e+00  2.280e-03 0.003
##
## Residual standard error: 0.3988 on 110 degrees of freedom
## Multiple R-squared:  0.7564, Adjusted R-squared:  0.7409
## F-statistic: 48.79 on 7 and 110 DF,  p-value: < 2.2e-16
```

Сравним ее с моделями 1, 2, 6:

```
## J test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + people:I(1 - pump_2) +
##     ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter + pump_1
## Model 2: log_consum ~ ave_temp + HDD + CDD + people + counter + pump_1 +
##     pump_2
##                  Estimate Std. Error t value  Pr(>|t|)
## M1 + fitted(M2) -0.15275     0.54552 -0.2800      0.78
## M2 + fitted(M1)  1.05996     0.22753  4.6586 9.047e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## J test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + people:I(1 - pump_2) +
##     ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter + pump_1
```

```
## Model 2: log_consum ~ counter + pump_1 + pump_2
##                 Estimate Std. Error t value Pr(>|t|)
## M1 + fitted(M2) -0.72738    0.89993 -0.8083   0.4207
## M2 + fitted(M1)  1.00742    0.08793 11.4573   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## J test
##
## Model 1: log_consum ~ HDD:people + people:I(counter) + people:I(1 - pump_2) +
##     ave_temp:I(1 - counter) + CDD:I(1 - pump_1) + counter + pump_1
## Model 2: log_consum ~ HDD:people + people:I(counter) + ave_temp:I(1 -
##     counter) + CDD:I(1 - pump_1) + counter + pump_1 + pump_2
##                 Estimate Std. Error t value Pr(>|t|)
## M1 + fitted(M2) -0.64031     0.7922 -0.8083  0.42070
## M2 + fitted(M1)  1.58863     0.7475  2.1253  0.03582 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Новая модель лучше.

Получившиеся уравнение линейной регрессии :

$y = -1.435 * counter - 5.867 * 10^{-1} * pump_1 + 1.26 * 10^{-4} * HDD * people + 6.248 * 10^{-1} people * counter + 2.56 * 10^{-1} * peo$

Расстояние Кука

Регрессия сильно подстраивается под далеко стоящие наблюдения.
Расстояние Кука — мера воздействия i-го наблюдения на регрессионноеуравнение:

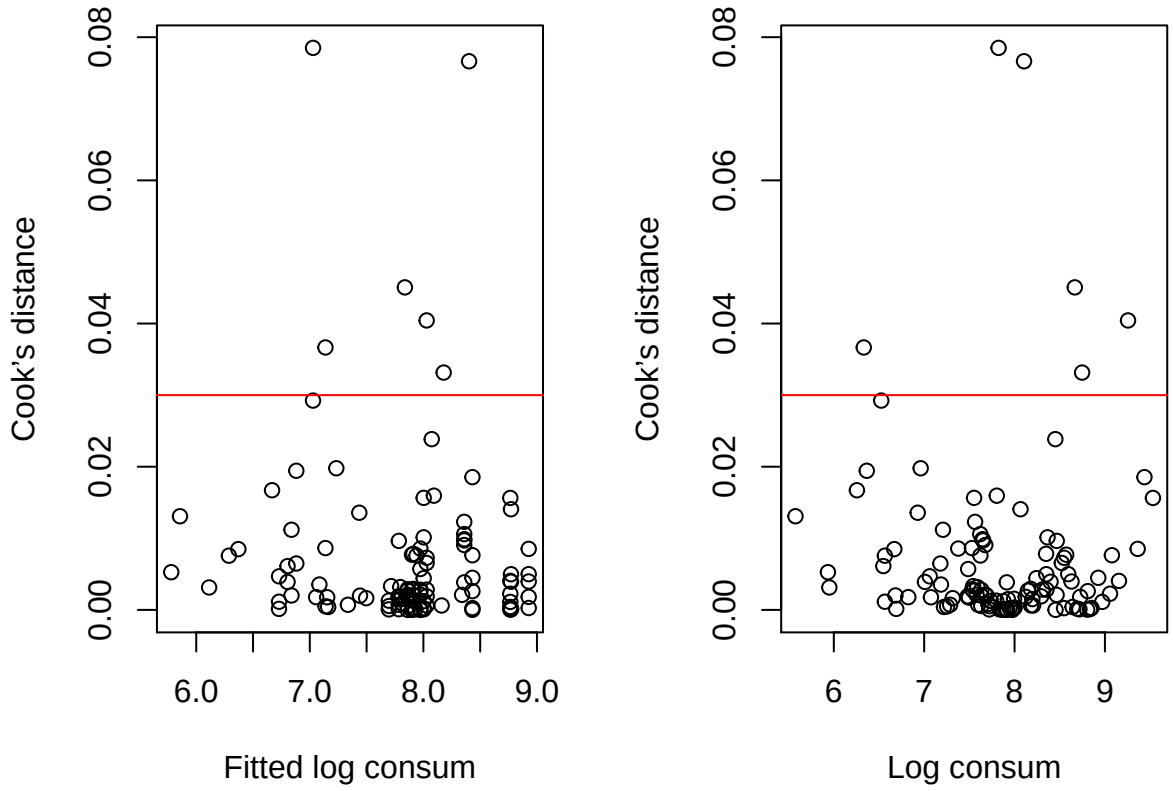
Расстояние Кука измеряет «расстояние» между B_j и B_{j-i} путем вычисления F-теста для гипотезы, что $B_j = B_{j-i}$, для J = 0,1 , ..., к. F -статистика вычисляется для каждого наблюдения следующим образом:

$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

$$(D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{RSS(k+1)} = \frac{\hat{\varepsilon}_i^2}{RSS(k+1)} \frac{h_i}{(1-h_i)^2})$$

$y\hat{}_j(i)$ — предсказания модели, настроенной по наблюдениям
1, . . . , i - 1, i + 1, . . . , n, для наблюдения j;
h_i — диагональный элемент матрицы ($H = X(X^T X)^{-1} X^T$)

Посмотрим на влиятельные наблюдения:



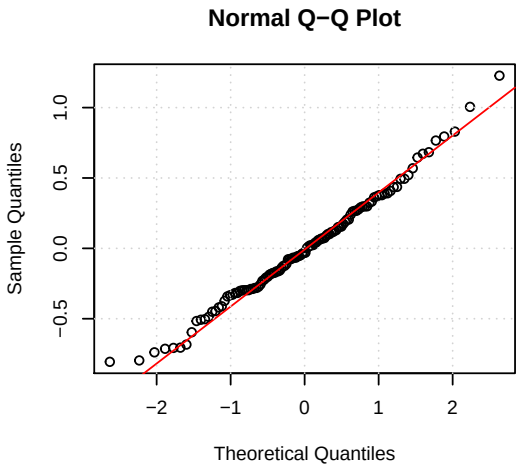
Удалим наблюдения с расстоянием Кука больше 0.03 (порог выбран визуально) и перенастроим модель 7.

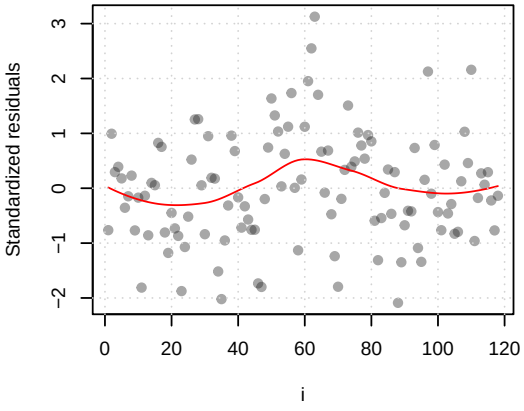
Сравним коэффициенты новой модели и модели 8:

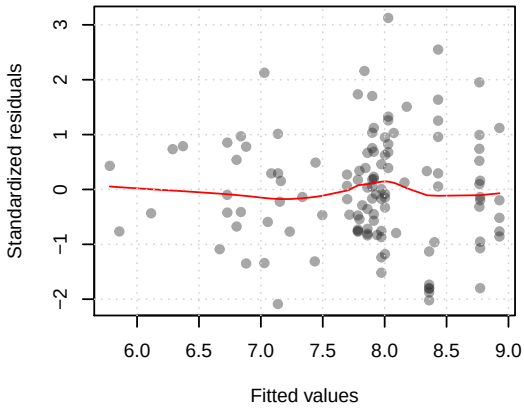
```
##                                All data Filtered data
## (Intercept)                7.8436040425  7.5259094388
## counter                   -1.4352131140 -1.3249912849
## pump_1                    -0.5867455534 -0.5569412628
## HDD:people                 0.0001259522  0.0001438199
## people:I(counter)          0.6247970444  0.6671996865
## people:I(1 - pump_2)       0.2560416919  0.2791894829
## ave_temp:I(1 - counter) -0.0192309743 -0.0158161953
## CDD:I(1 - pump_1)          0.0018649777  0.0017534103
```

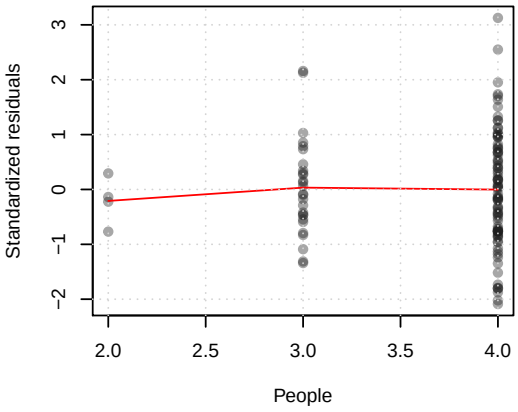
Коэффициенты изменились незначительно, поэтому не будем удалять влиятельные наблюдения.

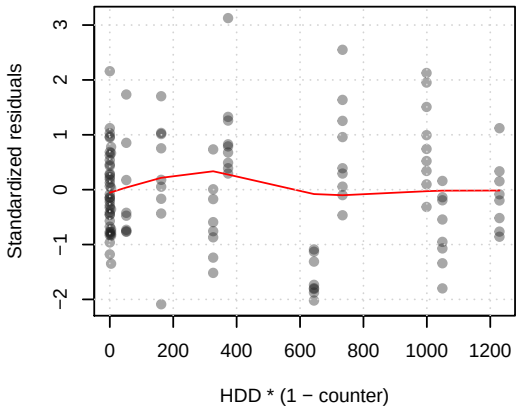
Визуальный анализ остатков:

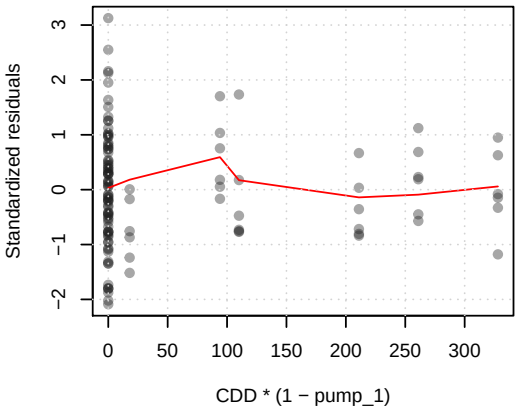

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles

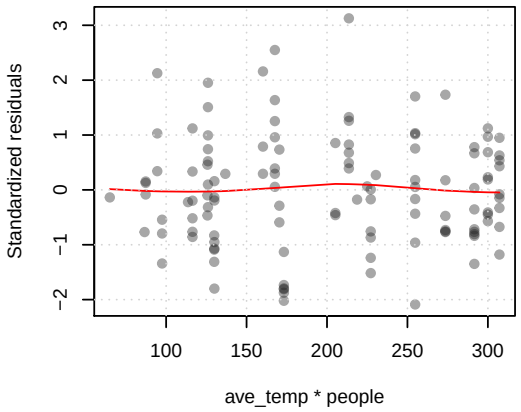

Standardized residuals
i


Standardized residuals
Fitted values


Standardized residuals
People

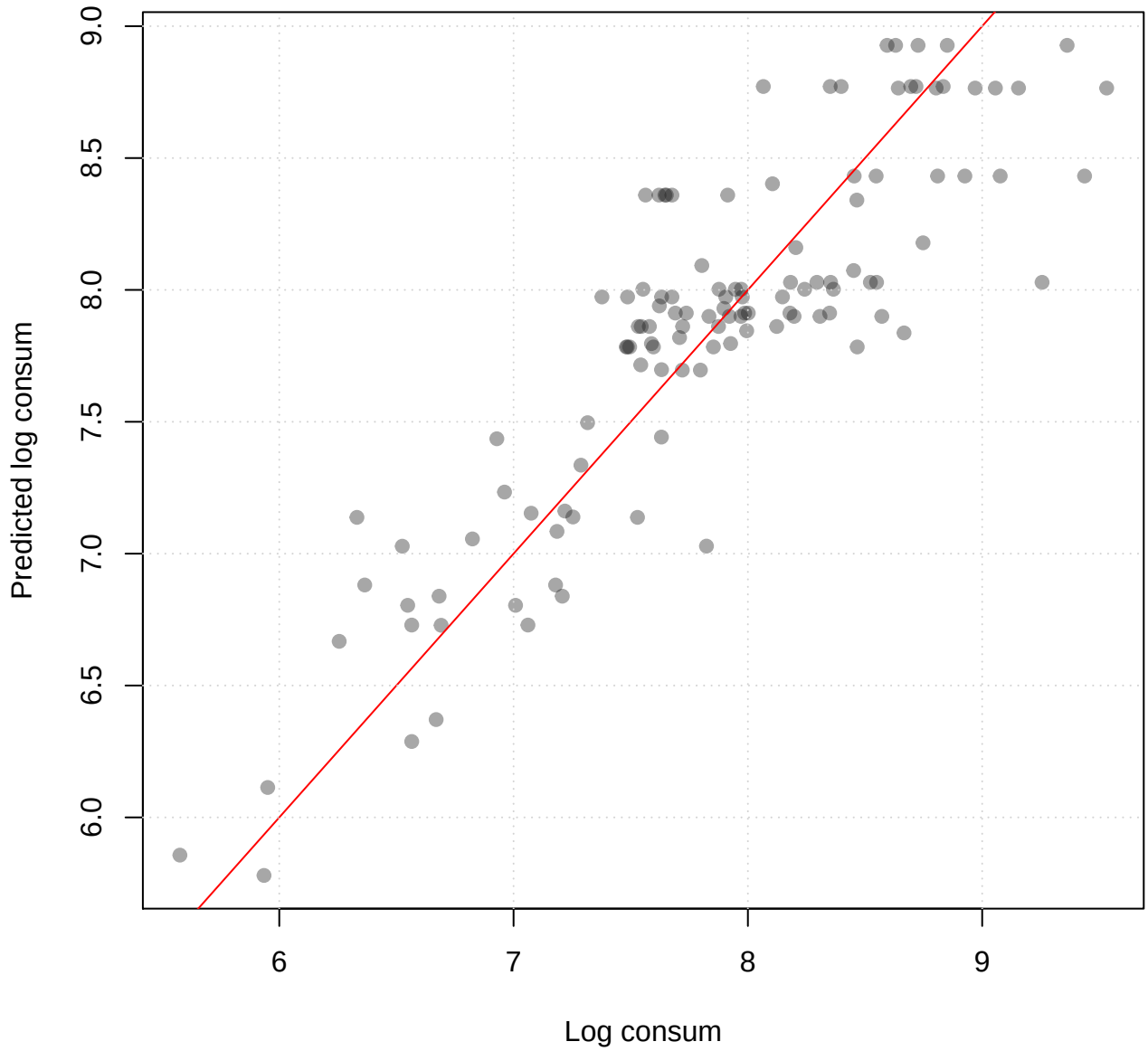

Standardized residuals
HDD * (1 − counter)


Standardized residuals
CDD * (1 − pump_1)


Standardized residuals
ave_temp * people

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 дополнительных факторов

- установка нового счетчика уменьшила потребление в 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%.

5. Заключение

В данной работе была рассмотрена проблема постороения оптимальной модели линейной регрессии. Подход, предлагаемый для ее решения - шаговая регрессия, подразумевает отбор из большого количества субъектов небольшой подгруппы переменных, которые вносят наибольший вклад в вариацию зависимой переменной. Одной из основных проблем, с пошаговой регрессии является то, что он ищет большое пространство возможных моделей. Поэтому он склонен к переобучению данных. Другими словами, пошаговая регрессия показывает лучшие результаты на подготовленной выборке, нежели на новых данных. Эта проблема может быть уменьшена, если критерий для добавления (или удаление) переменной является достаточно жестким. Моментом, в который надо остановиться и подумать, потому что следующие решения предопределены и ничего нельзя с этим сделать.

В курсовой работе, после каждой постоенной модели мы проверяли остатки на нормальность и гетероскедастичность, и далее исправляли это применяя различные критерии, например, критерийй Бокса-Кокса, для данных, которые не проходили тест на нормальность.

Таким образом, после построения оптимальной модели линейной регресси и ее визуализации, можно сделать выводы о там как скажется установка нового оборудования на дальнейшем потреблении.

Список литературы

[1] : Chong I.-G. Perfomance of some variable selection methods when multicollinearity is present.
[2] : Курс лекций Рябенко Евгений “Прикладной статистический анализ данных”, 2016.
[3] : Дрейпер Н.Р., Смит Г. Прикладной регрессионный анализ. — М.: Издательскийдом «Вильямс», 2007.
[4] : Кобзарь А.И. Прикладная математическая статистика. — М.: Физматлит, 2006.
[5] : Bretz F., Hothorn T., Westfall P. Multiple Comparisons Using R. — Boca Raton: Chapman and Hall/CRC, 2010.