Ласкин М.Б.

Модель множественной линейной регрессии и ее общая корректировка

30 мая 2024 г.

Общие положения

Из-за специфического распределения цен на рынке недвижимости
Модель вида (мультипликативная):

$$ln\ (V) = a_0 + a_1 f(x_1) + a_2 f(x_2) + a_3 f(x_3) + \ldots + a_n f(x_n) + \varepsilon$$

всегда лучше, чем аддитивная модель

$$V = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n + \varepsilon$$

Поэтому модель множественной линейной регрессии лучше строить для логарифмов. Почему? Потому что по условиям теоремы Гаусса-Маркова ошибки должны быть симметричны, гомоскедастичны, независимы и иметь нулевое математическое ожидание. Нормальность не требуется, но является очевидным бонусом.

Общие положения

Модель в векторном виде

$$E(y) = f(\bar{x}) = c_0 \times x_0 + c_1 \times x_1 + c_2 \times x_2 + \cdots + c_n \times x_n = (\bar{c}, \bar{x})$$

$\bar{c}$- вектор коэффициентов, $\bar{x}$- случайный вектор значений ЦОФ, $x_0 = 1$.
Имеется m наблюдений (объектов сравнения) $y_i, i = 1, m$ и n ЦОФ.
Матричная запись:

$$X = \begin{pmatrix} 1 & x_{11} \dots & x_{1n} \\ 1 & x_{21} \dots & x_{2n} \\ 1 & x_{m1} \dots & x_{mn} \end{pmatrix}$$

Целевая функция $\sum_{i=1}^{n} \left(y_i - (\bar{c} \times \bar{x_i})\right)^2 \to min$

или в матричной форме

$$RSS = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \left(Y - X\hat{C}\right)^T \left(Y - X\widehat{C}\right) \to min$$

$$\frac{\partial RSS}{\partial \hat{C}} = -2X^T Y + 2X^T X\hat{C} = 0$$

И решение (оценка коэффициентов)

$$\hat{C} = (X^T X)^{-1} X^T Y$$

Вывод 1: в матрице *X не* должно быть линейно зависимых столбцов (ЦОФ) и строк.

Вывод 2: после проведения корректировок модель множественной линейной регрессии применятся НЕ МОЖЕТ.

Множественная линейная регрессия как метод машинного обучения

Базовый принцип машинного обучения.

Исходное множество данных (датасет) делится на два или больше подмножеств: обучающее множество, тестовое множество. При необходимости дополнительных преобразований могут вводится верификационные множества. Конечная цель – создать модель машинного обучения, способную предсказывать значения целевой функции для наборов ЦОФ, для которых истинное значение неизвестно.

Здесь мы создадим из исходного набора данных:

- обучающее множество (train)

- валидационное множество (validation)

- тестовое множество (test)

Множественная линейная регрессия как метод машинного обучения

Исходные данные: ГБУ «Ленкадоценка».

3390 свободных земельных участков с назначением ИЖС в Ленинградской области, выставлявшихся на продажу в 2022 г. (данные для кадастровой оценки на 01.01.2023 г.

- обучающее множество (train) – 2400
- валидационное множество (validation) - 400
- тестовое множество (test) - 496

Множественная линейная регрессия как метод машинного обучения

На обучающем множестве получена модель:

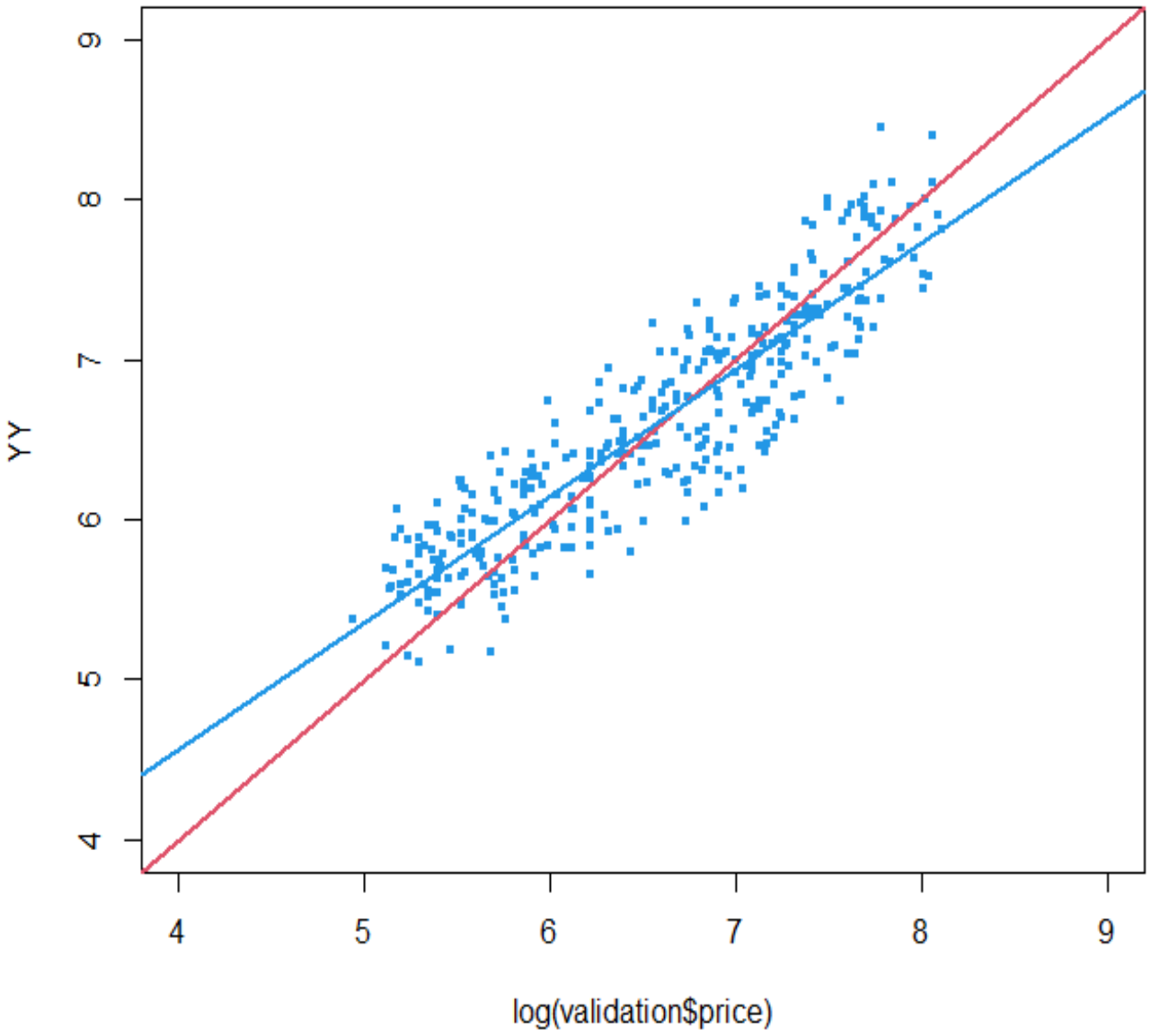
```
Call:
lm(formula = log(price) ~ log(S) + power + water + gaz + DistFedRoad +
    log(DistSpb) + DistRegCenter + DistTBO + Population + local.industry +
    Shop + School, data = xx)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91302 -0.24626 -0.01644  0.24223  0.90049

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.4443579  0.1282913  81.411  < 2e-16 ***
log(S)         -0.2341159  0.0165081 -14.182  < 2e-16 ***
power          -0.0575971  0.0407894  -1.412  0.15806
water           0.1059642  0.0201207   5.266 1.52e-07 ***
gaz             0.1476506  0.0206042   7.166 1.02e-12 ***
DistFedRoad    -0.0053472  0.0012259  -4.362 1.34e-05 ***
log(DistSpb)   -0.7552367  0.0121290 -62.267  < 2e-16 ***
DistRegCenter   0.0121003  0.0003967  30.504  < 2e-16 ***
DistTBO         0.0039883  0.0007825   5.097 3.72e-07 ***
Population      0.1504973  0.0098171  15.330  < 2e-16 ***
local.industry  0.0775871  0.0252427   3.074  0.00214 **
Shop            0.1664413  0.0191666   8.684  < 2e-16 ***
School         -0.1281480  0.0241839  -5.299 1.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3538 on 2387 degrees of freedom
Multiple R-squared:  0.8066,    Adjusted R-squared:  0.8057
F-statistic: 829.8 on 12 and 2387 DF,  p-value: < 2.2e-16
```

Множественная линейная регрессия как метод машинного обучения



На валидационном множестве – проверка.

По горизонтали - истинные значения,

по вертикали - предсказанные значения.

Красная линия – линия точных предсказаний,

синяя линия – тренд.

Стандартное отклонение предсказаний sd=0.383

Мы бы хотели, чтобы точки отклонялись от биссектрисы на столько, на сколько они отстоят от тренда.

Пусть произвольная точка имеет координаты (x^*, y^*).

Уравнение линии точных предсказаний (биссектрисы) $y = x$

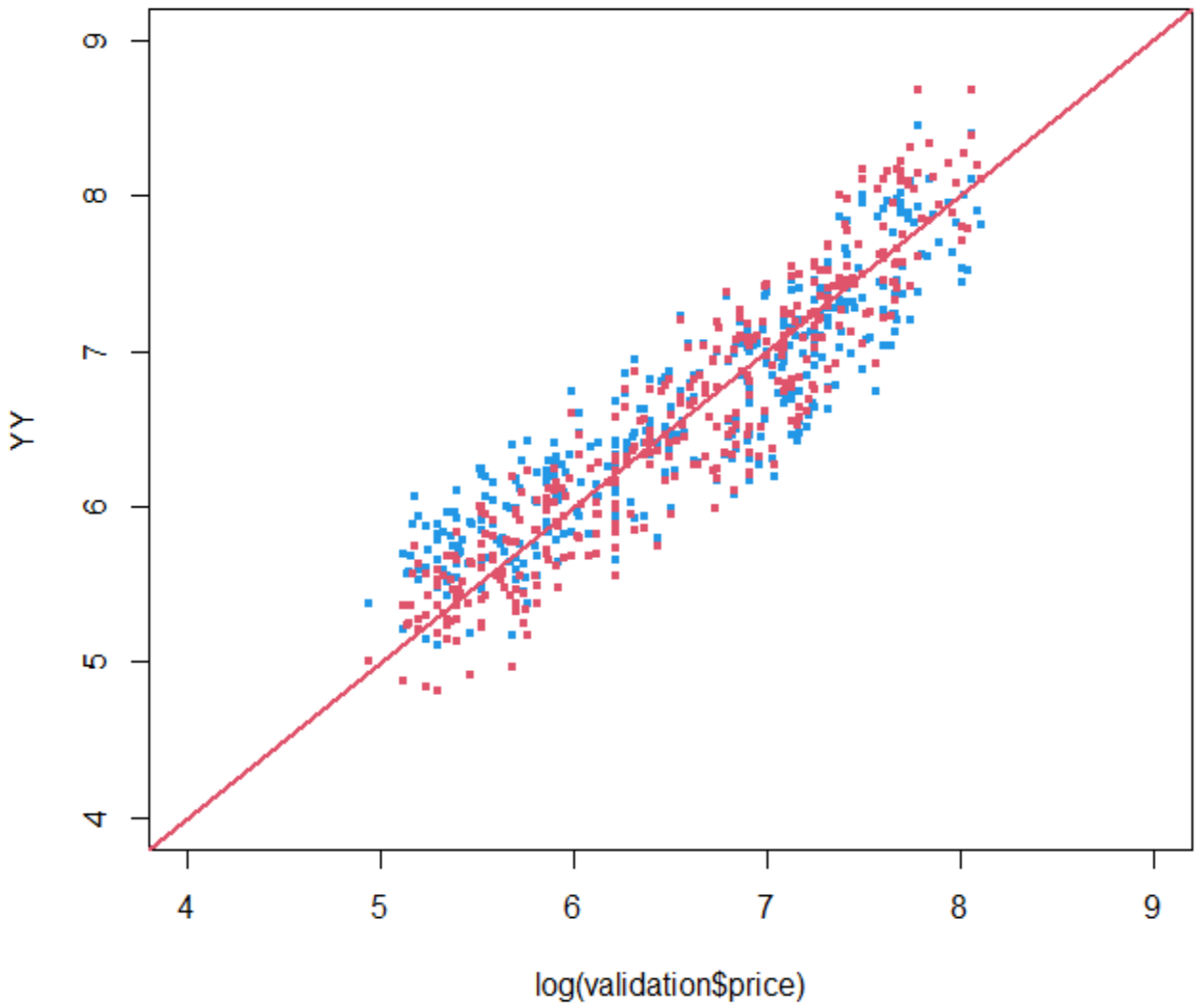
Уравнение линии тренда $y = k \cdot x + b$

Расстояние от произвольной точки до тренда $y^* - k \cdot x^* - b$

Скорректированное значение предсказания $x^* + y^* - k \cdot x^* - b$

Этого достаточно, чтобы проверить корректировку на тестовом множестве, но не достаточно для предсказаний, т.к. мы не знаем истинного значения (в отличие от тестового множества, для которого нам истинное значение известно.

Множественная линейная регрессия как метод машинного обучения



На рисунке результат корректировки.

Стандартное отклонение остается высоким.

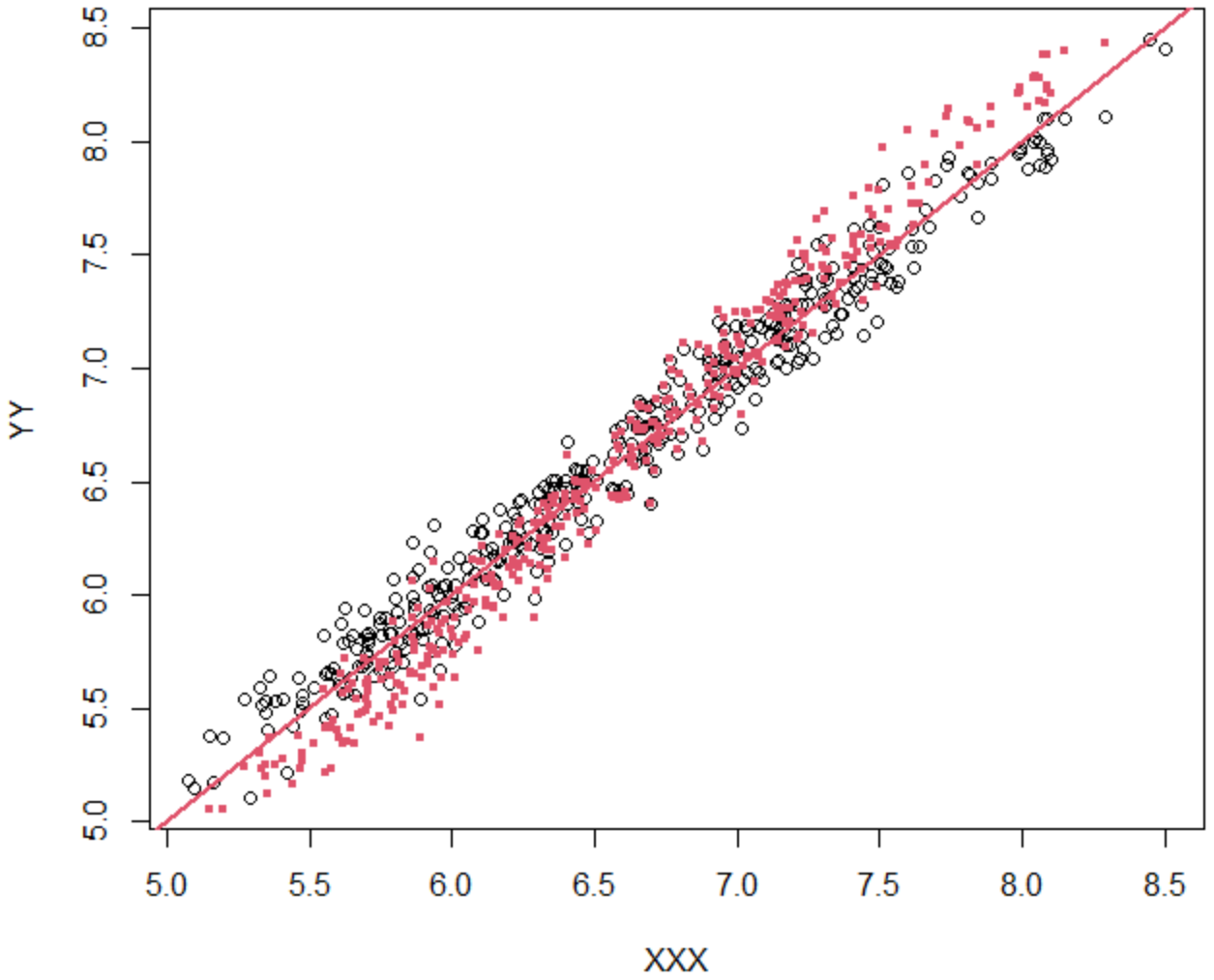
Примененная корректировка не позволяет распространить результат на множество для которого мы не знаем истинных значений

Но мы уже знаем структуру множества и полагаем, что она соответствует структуре подобных точек в выбранном секторе недвижимости

Поэтому мы знаем, что при фиксированном *y* мы можем ожидать характерное распределение возможных истинных значений

Сгенерируем случайным образом под каждое предсказание возможное истинное значение и посмотрим, что получится

Множественная линейная регрессия как метод машинного обучения



На рисунке черные точки сгенерированные возможные истинные значения против предсказанных

Красные точки – результат корректировки по тому же правилу, что и ранее, только теперь вместо настоящих истинных значений по горизонтали стоят значения, разыгранные с помощью случайного генератора.

Такую процедуру мы можем повторить для любого датасета, в котором истинных значений нет.

Точность предсказаний – управляется стандартным отклонением.

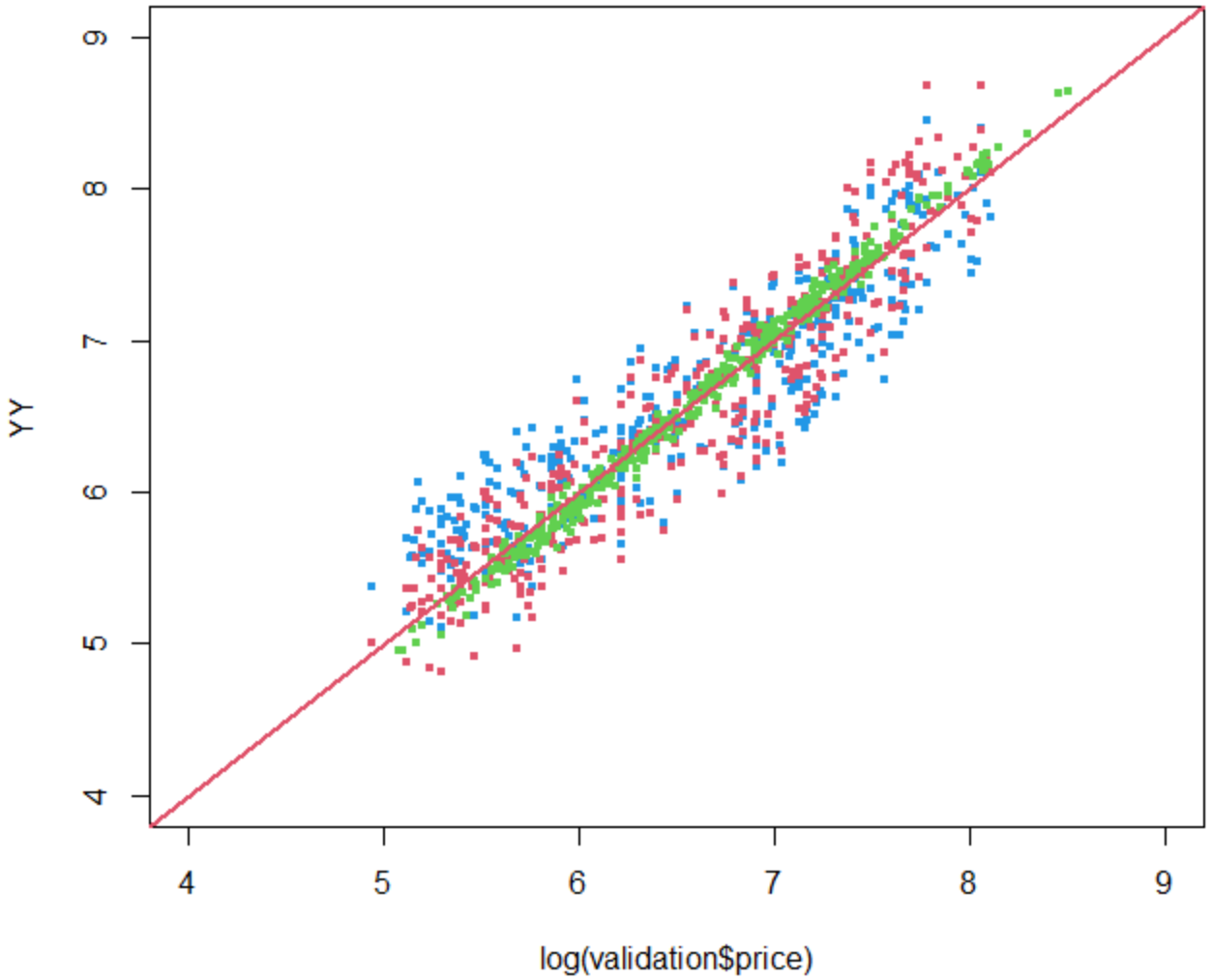
В формулу для скорректированных значений $x^* + y^* - k \cdot x^* - b$ добавляем коэффициент сжатия α

Новая формула для скорректированного предсказания

$$x^* + (y^* - k \cdot x^* - b) \cdot \alpha$$

Здесь x^* - сгенерированное возможное истинное значение, y^* - предсказание

Множественная линейная регрессия как метод машинного обучения



На рисунке голубые точки - истинные значения против предсказанных

Красные точки – результат корректировки

Зеленые точки – результат сжатия

Стандартное отклонение предсказаний, очевидно, снизилось

Спасибо за внимание!