Linear Regression. Линейная регрессия.

Data Science.

Lectures. Week 1.

Polina Kravets
20 сентября 2022 г.

Содержание

1 Регрессионный анализ в эконометрике

Регрессионный анализ - исследование зависимости одной переменной от другой.

- Зависимая переменная - переменная, изменение которой хотим объяснить.
- Переменные, с помощью которых объясняем эти изменения, называются независимыми, или факторами.

Облако рассеивания - инструмент, с помощью которого можно оценить вид зависимости.

Простая точечная диаграмма для двух переменных:

- y - зависимая переменная, вертикальная ось;
- x - независимая переменная, горизонтальная ось.

Каждая точка представляет собой пару наблюдений (x;y).

Проведенная прямая - попытка оценить, является ли зависимость (x;y) линейной; насколько далеко ложатся точки от данной прямой.

Визуально по облаку рассеивания можно оценить зависимость и вид зависимости.

Если имеет место линейная зависимость (x;y), то облако будет сильно вытянутым. Если облако имеет круглую форму - линейная зависимость отсутствует.

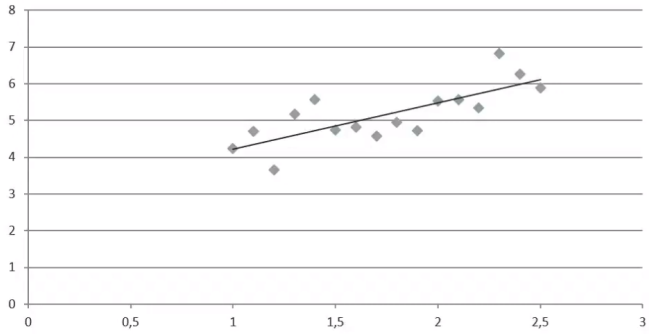

Рис. 1: Пример облака рассеивания

- В зависимости от того, как располагается прямая, близко к которой располагаются точки, можно понять, какому изменению переменной y будет соответствовать изменение переменной x;
- Прямая направлена вверх → рост одной переменной соответствует росту другой переменной; вниз → противоположно;
- ***Важно помнить, что этот визуальный инструмент не дает никаких математических вычислений и позволяет лишь выдвигать гипотезы о зависимости y от x.***

2 Уравнение регрессии

Математически уравнение регрессии, выражающее зависимость y от x, представляет собой:

$$Y_i = f(T, X_i, \varepsilon_i)$$

где Y_i - объясняемая переменная, X_i - объясняющая переменная (фактор), f - определяет модель регрессии, T - набор параметров модели, ε_i- случайные ошибки.

Если функция f линейна, то соответствующая регрессия называется линейной.

Это можно записать в виде условного математического ожидания:

$$E(Y_i|X_i) = B_0 + B_1 X_i$$

где B_0 - y-пересечение (intercept-term), B_1 - коэффициент наклона.

Без использования условного математического ожидания:

$$Y_i = B_0 + B_1 X_i + \varepsilon_i$$

Нужно выбрать модель регрессии так, чтобы она адекватно описывала зависимость y от x.

ε_i - независимая случайная величина с нулевым математическим ожиданием, имеет нормальное распределение.

2.1 Выборочное уравнение в линейной регрессии

На практике имеется реализация выборки, т.е. точные значения коэффициентов B_0 и B_1 неизвестны.

Выборочное уравнение линейной регрессии:

$$Y_i = b_1 X_i + b_0 + e_i$$

b_0, b_1 отличаются от истинных значений, $e_i = Y_i - b_0 - b_1 X_i$ - остатки, $e_i \neq \varepsilon_i$.

2.2 Свойства линейной регрессии

- Независимая переменная входит в уравнение как есть, без преобразований;
- Уравнение регрессии линейно по отношению к коэффициентам модели;
- Линейная регрессия покрывает множество случаев нелинейной - с помощью преобразования данных можно свести нелинейную регрессию к линейной. Например, взяв логарифм выражения $Y_i = e^{b_0} X_i^{b_1} e^{e_i}$, получим $\ln(Y_i) = b_0 + b_1 \ln(X_i) + e_i$, тем самым преобразовав нелинейную регрессию в линейную относительно логарифмов.

3 Метод наименьших квадратов

Суть метода: Из Y_i вычитаем $b_1 X_i - b_0$, т.е. берем остатки e_i, возводим их в квадрат и суммируем по всем i.

Метод наименьших квадратов (МНК) состоит в том, что мы находим такие значения b_0 и b_1, чтобы сумма квадратов остатков была наименьшей:

$$\Sigma_i e_i^2 = \Sigma_i (Y_i - (b_1 X_i + b_0))^2$$

$$b_1 = \frac{(\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2} = \frac{Cov(X, Y)}{Var(X)}$$

b_0 - точка пересечения регрессионной прямой с осью У при X=0.

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Регрессионная прямая $Y = b_1 X + \bar{Y} - b_1 \bar{X}$ всегда проходит через точку с координатами $(\bar{X}; \bar{Y})$.

3.1 Основные предположения для использования МНК

Ключевые предположения:

- Y_i являются независимыми одинаково распределенными случайными величинами. Если X_i тоже, то они должны быть независимо одинаково распределены;
- $E(\varepsilon_i | X_i) = 0$;
- В выборке нет выбросов;

Дополнительные предположения:

- Зависимая переменная линейным образом зависит от независимой;
- Изменение независимой переменной объясняется только зависимой;
- Если X_i - случайные величины, то они не должны зависеть от ε_i;
- Все ошибки независимы между собой;
- Дисперсия ε_i одинакова, ε_i имеют нормальное распределение;

3.2 Преимущества использования МНК

- Оценки коэффициентов являются несмещенными, состоятельными, эффективными;
- Для больших объемов наблюдений оценки b_0 и b_1 являются асимптотически нормальными;
- В случае двух переменных, оценки коэффициентов легко считаются;
- Подсчет коэффициентов и интерпретация и анализ результатов легко понимаются среди множества различных сфер;

4 Интерпретация результатов линейной регрессии и МНК-оценки

Насколько модель адекватно описывает зависимость?

- **Сумма квадратов остатков (SSR)**: вычисляется как Σe_i^2;
- Чем меньше SSR, тем лучше ложатся точки на регрессионную прямую, тем более адекватно модель описывает зависимость;
- Минус показателя: размерная величина.
- **Коэффициент детерминации (R^2)**:

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(\bar{Y}_i - \bar{Y})^2 + \Sigma(Y_i - \bar{Y}_i)^2$$

$\Sigma(\bar{Y}_i - \bar{Y})^2$ - объясненная сумма квадратов (ESS),
$\Sigma(Y_i - \bar{Y}_i)^2$ - сумма квадратов остатков (SSR),
$\Sigma(Y_i - \bar{Y})^2$ - общая сумма квадратов (TSS).

$$R^2 = \frac{ESS}{TSS} = \frac{\Sigma(\bar{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} = 1 - \frac{\Sigma(Y_i - \bar{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2}$$

- Коэффициент детерминации всегда принимает значения в диапазоне [0,1]; чем ближе к 1, тем лучше точки ложатся на прямой;
- Физический смысл величины: $|r| = \sqrt{R^2}$ - модуль коэффициента корреляции.

5 Стандартная ошибка регрессии

Стандартная ошибка регрессии (SER) используется при построении доверительных интервалов, при оценки точности прогнозов:

$$SER = \sqrt{\frac{\Sigma e_i^2}{n-2}}$$

Стандартная ошибка регрессии показывает, насколько хорошо точки ложатся на прямую. Чем меньше SER, тем лучше точки ложатся на прямую.

6 Вычисление доверительных интервалов для коэффициентов регрессии

Доверительный интервал для коэффициента b_1:

$$b_1 + \pm t_c s_{b1}$$

t_c - критическое двустороннее значение, полученное из таблицы Стьюдента с числом степеней свободы $n-2$,
s_{b1} - стандартная ошибка коэффициента b_1

$$s_{b1} = \frac{\sqrt{\frac{1}{n-2}\Sigma e_i^2}}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

7 Тестирование гипотез для коэффициентов

- Нулевая гипотеза: $B_1 = B$ (какому-то значению), альтернатива $B_1 \neq B$;
- Применяем критерий Стьюдента:
$$t = \frac{b_1 - B}{s_{b1}}$$
- Если $t > +t_{\text{critical}}$ или $t < -t_{\text{critical}}$, нулевая гипотеза отвергается;
- Можно использовать p-значение - наименьший уровень значимости, при котором нулевая гипотеза может быть отвергнута;
- Для двустороннего критерия p-значение будет в два раза больше, чем для односторонней выборки;

8 Понятие статистической значимости

- **Статистическая значимость** - проверка гипотезы о том, что какой-то коэффициент равен 0;
- Величина называется статистически значимой, если гипотеза о том, что она равна 0, отвергается на данном уровне значимости;
- Для уравнения регрессии есть смысл тестировать коэффициент b_1: если коэффициент становится статистически незначимым, это означает, что y_i не зависит от x, по крайней мере, линейным образом;

9 Прогноз

- **Прогноз** в уравнении регрессии для переменной Y в точке X_p:
$$\bar{Y} = b_1 X_p + b_0$$
- Если выполняются все предположения линейной регрессии МНК, то можно построить доверительный интервал для прогноза:
$$\bar{Y} \pm t_c s_f$$
t_c - критическое значение двустороннего распределения Стьюдента для заданного уровня значимости и $n-2$ степенями свободы,
s_f - стандартная ошибка прогноза:
$$s_f^2 = SER^2(1 + \frac{1}{n} + \frac{(x_p - \bar{X}^2)}{(n-1)s_x^2})$$

10 Фиктивные переменные

- **Фиктивные переменные** - независимые переменные, которые принимают два значения 0 (если событие не произошло) и 1 (если событие произошло);
- Используются, когда независимая переменная является изначально бинарной;
- Рассчитанный коэффициент регрессии для фиктивных переменных показывает разницу между зависимой переменной для категории, представленной этой фиктивной переменной, и зависимой переменной для всех классов за исключением класса фиктивной переменной;

11 Гомоскедастичность и гетероскедастичность

- МНК применяется при ряде предположений, одно из которых о том, что остатки имеют нормальное распределение с математическим ожиданием 0;
  Этот случай называется **гомоскедастичностью.**
- Если предположение неверное, то говорят о **гетероскедастичности**;
- **Безусловная гетероскедастичность** означает, что гетероскедастичность не связана с величиной независимой переменной X.
  При большом объеме выборки все выводы остаются адекватными.
- **Условная гетероскедастичность** означает, что дисперсия остатков зависит от величины независимой переменной x;
  В этом случае рассмотренные методы не применимы: оценки МНК уже не будут несмещенными, эффективными и состоятельными; оценки не будут иметь нормальное распределение и к ним не будут применимы доверительные интервалы и критерий Стьюдента;
- В случае гомоскедастичности на графике остатков диапазон по вертикали остается практически одним и тем же; в случае гетероскедастичности;

На графике можно выделить 2 группы.

Если возьмем x примерно меньше 1,7, то остатки достаточно близко к нулю, в другой группе - существенно больше нуля.

В левой группе дисперсия оказывается существенно меньше, чем в правой.

Получили визуальную оценку гетероскедастичности данных.

Чтобы избавиться от гетероскедастичности, можно попытаться воспользоваться альтернативными МНК, либо каким-то образом поработать с данными.

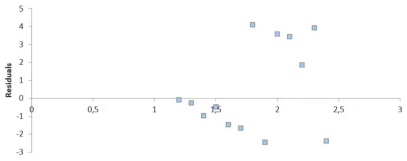

Рис. 2: Residual plot

12 Теорема Гаусса-Маркова

Теоремой Гаусса-Маркова называется следующее утверждение: если условия применимости метода наименьших квадратов выполнены, то МНК-оценки являются несмещенными, эффективными, состоятельными, асимптотически нормальными.

Что делать при невыполнении теоремы Гаусса-Маркова:

- взять не сумму наименьших квадратов, а сумму наименьших абсолютных отклонений;
- рассмотреть взвешенные наименьшие квадраты;

13 Маленькое число наблюдений

- В случае, когда число наблюдений велико, можно применить центральную предельную теорему: все доверительные интервалы и применение критериев обосновано. Важно, чтобы распределение ошибок было не важно каким, но не менялось;

- В случае малых объемов выборки (< 30) обязательно проверить, что остатки имеют нормальное распределение с одной и той же дисперсией;