

Анализ данных в политических науках. Линейная регрессия

Н. В. Артамонов, И. А. Истомин

МГИМО МИД России

9 декабря 2017 г.



Содержание

- 1 Стандартная линейная модель регрессии
 - Основные предположения
 - Интерпретация коэффициентов
- 2 Метод наименьших квадратов
- 3 Тестирование гипотез
- 4 Прогнозирование

Стандартная линейная модель регрессии CLRM (Classical Linear Regression Model)

- y – **зависимая переменная** (dependent variable);
- $x_1, \dots, x_k$ – **объясняющие переменные, влияющие переменные, регрессоры**.

По каждому из факторов имеем выборочные наблюдения

$$y_i, x_{i1}, \dots, x_{ik} \quad i = 1, \dots, n$$

Далее всегда

- n – объём выборки,
- k – число объясняющих переменных.

- 1 Стандартная линейная модель регрессии
 - Основные предположения
 - Интерпретация коэффициентов
- 2 Метод наименьших квадратов
- 3 Тестирование гипотез
- 4 Прогнозирование

Спецификация

Линейная модель регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \quad i = 1, \dots, n \quad (1)$$

где

- u_i – **ошибка** (error) модели регрессии.
- $\beta_0, \beta_1, \dots, \beta_k$ – **коэффициенты** регрессии (в генеральной совокупности).

Наблюдаемые факторы

Зависимая и объясняющие переменные:

- y – эндогенный **количественный** фактор;

Наблюдаемые факторы

Зависимая и объясняющие переменные:

- y – эндогенный **количественный** фактор;
- $x_1, \dots, x_k$ – экзогенные факторы, как **количественные**, так и **качественные**;

Наблюдаемые факторы

Зависимая и объясняющие переменные:

- y – эндогенный **количественный** фактор;
- $x_1, \dots, x_k$ – экзогенные факторы, как **количественные**, так и **качественные**;
- $y, x_1, \dots, x_k$ **наблюдаемы** и рассматриваются как случайные или детерминированные (регрессоры) величины.

Ошибка регрессии

Ошибка модели регрессии:

- u – случайная величина, моделирует влияние неучтённых факторов (вводится из-за недостатка информации);

Ошибка регрессии

Ошибка модели регрессии:

- u – случайная величина, моделирует влияние неучтённых факторов (вводится из-за недостатка информации);
- ошибка u **ненаблюдаема**.

Ошибка линейной регрессии

Условия на ошибку регрессии:

- ❶ Несистематичность (условия экзогенности объясняющих переменных)
- ❷ Однородность или гомоскедастичность: во всех наблюдениях “степень влияния” неучтённых факторов одинакова.
- ❸ Независимость наблюдений (отсутствие серийной корреляции)

Эти условия называются **условия Гаусса-Маркова**.

Ошибка линейной регрессии

Условия на ошибку регрессии:

- ❶ Несистематичность (условия экзогенности объясняющих переменных)
- ❷ Однородность или гомоскедастичность: во всех наблюдениях “степень влияния” неучтённых факторов одинакова.
- ❸ Независимость наблюдений (отсутствие серийной корреляции)

Эти условия называются **условия Гаусса-Маркова**.

Замечание

Второе условие часто “неадекватно” наблюдениям. Это, например, связано с “неоднородностью” данных.

- 1 Стандартная линейная модель регрессии
 - Основные предположения
 - Интерпретация коэффициентов
- 2 Метод наименьших квадратов
- 3 Тестирование гипотез
- 4 Прогнозирование

Коэффициенты регрессии

Коэффициенты $\beta_0, \beta_1, \dots, \beta_k$ линейной регрессии (1):

- показывают (**количественно**) как регрессоры влияют на зависимую переменную;
- a priori неизвестны, необходимо оценить по выборочным данным.

Коэффициенты регрессии. Интерпретация

Коэффициенты в линейной регрессии (1) при количественном факторе имеют смысл **средних предельных значений**:

Интерпретация коэффициентов

На сколько в среднем изменится зависимая переменная при **увеличении объясняющей переменной на единицу** (при прочих равных, *ceteris paribus*).

Замечание

Константа β_0 в общем случае **не интерпретируется**.

Коэффициенты регрессии. Интерпретация

Пример (Продолжительность сна)

sleep – недельная продолжительность сна (мин), *totwrk* – недельная занятость (мин), *age* – возраст.

Рассмотрим регрессию

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 age + u.$$

Как интерпретировать коэффициенты?

Качественные регрессоры

Для учёта качественных факторов используем **бинарные регрессоры** (dummy, binary variable).

Качественные регрессоры

Для учёта качественных факторов используем **бинарные регрессоры** (dummy, binary variable).

Пример

Гендерный фактор

$$gender = \begin{cases} 1 \\ 0 \end{cases}$$

Качественные регрессоры

Для учёта качественных факторов используем **бинарные регрессоры** (dummy, binary variable).

Пример

Гендерный фактор

$$gender = \begin{cases} 1 \\ 0 \end{cases}$$

Тогда коэффициент при бинарной переменной – “отдача” (в среднем) от “обладания” качественным признаком.

Качественные регрессоры

Пример (Зарплатное уравнение)

wage – зарплата, *age* – возраст, *male* – гендерный фактор.

Рассмотрим регрессию

$$wage = \beta_0 + \beta_1 age + \beta_2 male + u.$$

Тогда β_2 – средняя разница в оплате между М и Ж (измеряет “дискриминацию” по гендерному фактору).

Наряду с линейной регрессией рассмотрим регрессию

$$\ln y = \beta_0 + \beta_1 x_1 + \cdots + u$$

Наряду с линейной регрессией рассмотрим регрессию

$$\ln y = \beta_0 + \beta_1 x_1 + \dots + u$$

Интерпретация коэффициентов

При увеличении фактора x_1 **на единицу** (при прочих равных, *ceteris paribus*), зависимая переменная y **в среднем** **изменяется на $\beta_1 \cdot 100\%$** .

Рассмотрим также регрессию

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \cdots + u$$

Рассмотрим также регрессию

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \cdots + u$$

Интерпретация коэффициентов

При увеличении фактора x_1 на **1%** (при прочих равных, *ceteris paribus*), зависимая переменная y **в среднем** **изменяется на $\beta_1\%$** .

Замечание

Коэффициент β_1 имеет смысл коэффициента эластичности.

- 1 Стандартная линейная модель регрессии
 - Основные предположения
 - Интерпретация коэффициентов
- 2 Метод наименьших квадратов
- 3 Тестирование гипотез
- 4 Прогнозирование

Как оценить модель?

Задача

Как оценить модель на статистических данных? Т.е. как оценить параметры модели (коэффициенты)?

Основной метод оценивания: Метод Наименьших Квадратов или OLS-метод (OLS = Ordinary Least Squares)

Почему?: при выполнении стандартных условия на ошибку регрессии OLS-оценки **эффективны или оптимальны**.

OLS: случай 2D

Регрессия с одной объясняющей переменной

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Имеем наблюдения $\{x_i, y_i\}_{i=1}^n$, их можно рассматривать как n точек на плоскости (график рассеяния).

OLS: случай 2D

Идея метода

Найти прямую $y = \beta_0 + \beta_1 x$ “**наименее отклоняющуюся**” от всех точек $\{(x_i, y_i)\}_{i=1}^n$ в смысле суммы квадратов (вертикальных) отклонений для каждой точки (“подогнать” прямую под данные):

$$SS = SS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

OLS: случай 2D

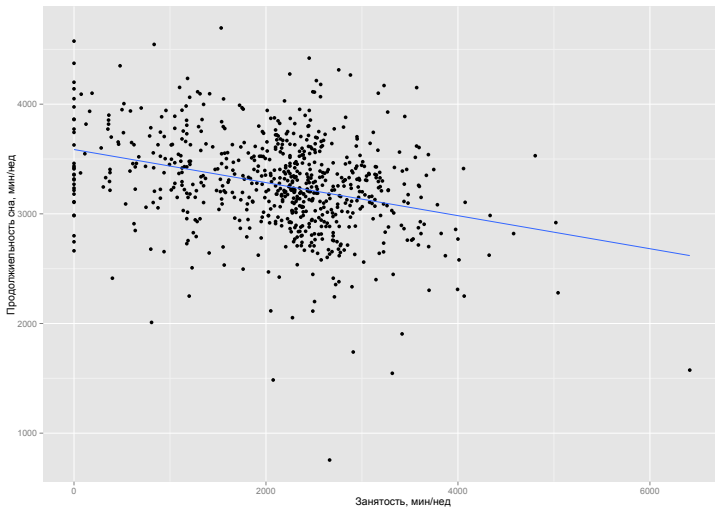
Идея метода

Найти прямую $y = \beta_0 + \beta_1 x$ “**наименее отклоняющуюся**” от всех точек $\{(x_i, y_i)\}_{i=1}^n$ в смысле суммы квадратов (вертикальных) отклонений для каждой точки (“подогнать” прямую под данные):

$$SS = SS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Тогда параметры оптимальной прямой есть решение оптимальной задачи (**при заданных** $\{x_i, y_i\}_{i=1}^n$)

$$\min_{\beta_0, \beta_1} SS = \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$



CLRM

oooooooooooooooo

OLS

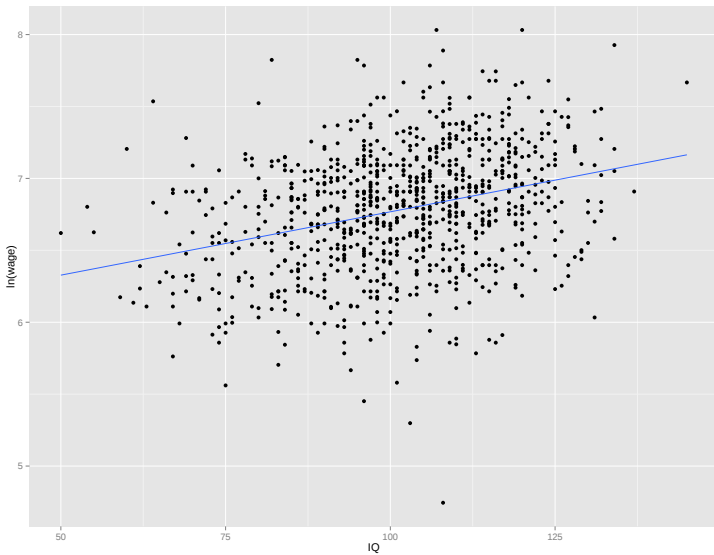
oooo●oooo

Inferences

oooooooooooo

Прогнозирование

ooo



OLS в общем случае

В общем случае OLS-метод нам даёт:

- 1 $\hat{\beta}_j$ – оценки коэффициентов
- 2 $s. e.(\beta_j)$ – стандартные ошибки коэффициентов
- 3 R^2 – показатель “качества подгонки” или коэффициент детерминации
- 4 R^2_{adj} – скорректированный или исправленный R^2 .

OLS в общем случае

В общем случае OLS-метод нам даёт:

- 1 $\hat{\beta}_j$ – оценки коэффициентов
- 2 $s.e.(\beta_j)$ – стандартные ошибки коэффициентов
- 3 R^2 – показатель “качества подгонки” или коэффициент детерминации
- 4 R^2_{adj} – скорректированный или исправленный R^2 .

$s.e.(\beta_j)$ & R^2 используются для тестирования гипотез

$R^2_{adj} = \bar{R}^2$ используется для сравнения разных регрессий

“Качество подгонки”

Важно!

Для каждой регрессии, наряду с оценками коэффициентов, вычисляется (относительный) **показатель “качества подгонки”** модели под данные со значениями от 0 до 1.

Для оценённой регрессии определим

- 1 $\hat{y}_i$ – подогнанные или предсказанные значения
- 2 $e_i = y_i - \hat{y}_i$ – остатки

Коэффициент R^2

Из определения

- $R^2 = 0 \iff \hat{y}_i = \bar{y} \iff \hat{\beta}_1 = \dots = \hat{\beta}_k = 0$ (“плохая подгонка” регрессии под данные, модель “ничего не объясняет”).
- $R^2 = 1 \iff y_i = \hat{y}_i, (i = 1, \dots, n)$ (“идеальная подгонка” регрессии под данные)

Коэффициент R^2

Из определения

- $R^2 = 0 \iff \hat{y}_i = \bar{y} \iff \hat{\beta}_1 = \dots = \hat{\beta}_k = 0$ (“плохая подгонка” регрессии под данные, модель “ничего не объясняет”).
- $R^2 = 1 \iff y_i = \hat{y}_i, (i = 1, \dots, n)$ (“идеальная подгонка” регрессии под данные)

Интерпретация R^2

Какая доля изменения зависимой переменной (на выборочных данных) определяется объясняющими переменными (объясняется регрессией).

Скорректированный коэффициент R^2

Свойства:

- $R^2_{adj} \leq R^2$
- м.б. отрицательным
- не интерпретируется!

Скорректированный коэффициент R^2

Свойства:

- $R^2_{adj} \leq R^2$
- м.б. отрицательным
- не интерпретируется!

Использование: сравнение моделей с одинаковой зависимой переменной (но с разными наборами регрессоров), критерий $\max \bar{R}^2$.

- 1 Стандартная линейная модель регрессии
 - Основные предположения
 - Интерпретация коэффициентов
- 2 Метод наименьших квадратов
- 3 Тестирование гипотез
- 4 Прогнозирование

Inferences: значимость коэффициента

Проверка значимости коэффициента регрессии: на данных тестируем («простую») гипотезу

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

Смысл

Тестируем гипотезу, что **в генеральной совокупности фактор x_j не влияет на зависимую переменную.**

Inferences: значимость коэффициента

На основе статистических данных нужно

- либо отвергнуть нулевую гипотезу (**коэффициент значим**, данные противоречат гипотезе).
- либо не отвергать нулевую гипотезу (**коэффициент незначим**, данные не противоречат гипотезе)

Inferences: t-тест для коэффициентов

Тестовая t-статистика:

$$t = \frac{\hat{\beta}_j}{\text{s. e.}(\beta_j)}$$

где $\hat{\beta}_j$ – оценка коэффициента, $\text{s. e.}(\beta_j)$ – стандартная ошибка коэффициента (всё автоматически вычисляется программами).

Inferences: t-тест для коэффициентов

Тестовая t-статистика:

$$t = \frac{\hat{\beta}_j}{s. e.(\beta_j)}$$

где $\hat{\beta}_j$ – оценка коэффициента, $s. e.(\beta_j)$ – стандартная ошибка коэффициента (всё автоматически вычисляется программами).

Критическое значение: распределения Стьюдента или t-распределения

$$t_{cr} = t_{n-k-1}(\alpha),$$

$(df = n - k - 1)$ – число степеней свободы.

Inferences: t-тест для коэффициентов

Статистическое правило t-теста

- Отвергаем H_0 при $|t| > t_{cr}$, **коэффициент значим** (тестовая статистика «большая» по модулю).
- Не отвергаем H_0 при $|t| < t_{cr}$, **коэффициент незначим** (тестовая статистика «маленькая» по модулю).

Альтернативно: используем Р-значения

Inferences: значимость регрессии

Проверка значимости регрессии «в целом»: тестируем гипотезу

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

Смысл

Все объясняющие переменные **в совокупности не влияют** на зависимую переменную (регрессоры «ничего не объясняют»).

Inferences: F-тест для регрессии

Тестовая F-статистика для проверки значимости «в целом» (автоматически вычисляется программами)

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$

Критическое значение: распределения Фишера или F-распределения

$$F_{cr} = F_{k, n-k-1}(\alpha).$$

$(df_1 = k, df_2 = n - k - 1)$ – степени свободы.

Inferences: F-тест для регрессии

Статистическое правило

- Отвергаем H_0 при $F > F_{cr}$, регрессия «в целом» **значима**.
- Не отвергаем H_0 при $F < F_{cr}$, регрессия «в целом» **незначима**.

Неформальное статистическое правило

Нулевую гипотезу отвергаем, если **тестовая F-статистика большая** (относительно критического значения).

Альтернативно: используем Р-значение

Inferences: совместная значимость

Проверка совместной значимости нескольких коэффициентов: тестируем гипотезу

$$H_0 : \beta_1 = \dots = \beta_J = 0 \quad (1 < J < k)$$

(в генеральной совокупности первые J коэффициентов равны нулю).

Смысл

В генеральной совокупности объясняющие переменные $x_1, \dots, x_J$ **не влияют** на зависимую переменную.

Inferences: F-тест для совместной значимости

Для вычисления тестовой статистики нужно оценить две («вложенные») регрессии:

- «длинную» (со всеми факторами, unrestricted)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \text{error}, \quad R^2$$

- «короткую» (без учёта факторов из H_0 , restricted)

$$y = \beta_0 + \beta_{J+1} x_{J+1} + \cdots + \beta_k x_k + \text{error}, \quad R^2_{restr}$$

Очевидно $R^2_{restr} \leq R^2$

Inferences: F-тест для совместной значимости

Тестовая F-статистика

$$F = \frac{R^2 - R_{restr}^2}{1 - R^2} \cdot \frac{n - k - 1}{J}$$

n – объем выборки, k – общее число факторов, J – число коэффициентов, совместную значимость которых тестируем.

Inferences: F-тест для совместной значимости

Тестовая F-статистика

$$F = \frac{R^2 - R_{restr}^2}{1 - R^2} \cdot \frac{n - k - 1}{J}$$

n – объем выборки, k – общее число факторов, J – число коэффициентов, совместную значимость которых тестируем.

Критическое значение: распределения Фишера или F-распределения

$$F_{cr} = F_{J, n-k-1}(\alpha).$$

$(df_1 = J, df_2 = n - k - 1)$ – степени свободы.

Inferences: F-тест для совместной значимости

Статистическое правило

- Отвергаем H_0 при $F > F_{cr}$, **совместное влияние факторов значимо**;
- Не отвергаем H_0 при $F < F_{cr}$, **совместное влияние факторов незначимо**.

Неформальное статистическое правило

Нулевую гипотезу отвергаем, если **тестовая F-статистика большая** (относительно критического значения).

Альтернативно: используем Р-значение

- 1 Стандартная линейная модель регрессии
 - Основные предположения
 - Интерпретация коэффициентов
- 2 Метод наименьших квадратов
- 3 Тестирование гипотез
- 4 Прогнозирование

Прогнозирование

Пусть заданы значения регрессоров из рассматриваемой генеральной совокупности $x_1, \dots, x_k$.

Как получить “оптимальный” прогноз зависимой переменной?

Два вида прогноза:

- точечный $\hat{y}$,
- интервальный (с заданной доверительной вероятностью γ).

Точечный прогноз

Утверждение

Пусть для линейной модели регрессии (1) выполнены условия теоремы Гаусса-Маркова.

Тогда при заданных значениях объясняющих переменных из генеральной совокупности **оптимальный линейный прогноз** равен

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

Точечный прогноз

Для регрессии

$$\ln y = \beta_0 + \beta_1(\ln)x_1 + \cdots + \beta_k(\ln)x_k + \text{error}$$

Оптимальный прогноз

$$\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1(\ln)x_1 + \cdots + \hat{\beta}_k(\ln)x_k.$$

Откуда

$$\hat{y} = \exp \left(\widehat{\ln y} \right)$$