

Лекция 4: Модель множественной линейной регрессии

Введение: почему одной переменной мало?

На прошлых лекциях мы предсказывали зарплату (Y) только на основе образования (X). Но мы интуитивно понимаем, что на зарплату влияет и опыт, и отрасль, и пол. Если мы не будем их учитывать, наша оценка влияния образования будет **смещенной**.

Модель множественной линейной регрессии позволяет оценить зависимость одной переменной от *нескольких* факторов одновременно, "очистив" влияние каждого из них от влияния остальных.

1. Общий вид модели и интерпретация коэффициентов

Матричная форма записи модели:

$$y = X\beta + \varepsilon$$

где:

- y — вектор зависимой переменной размера $n \times 1$
- X — матрица регрессоров размера $n \times (k+1)$ (включая столбец единиц для константы)
- β — вектор параметров размера $(k+1) \times 1$
- ε — вектор случайных ошибок размера $n \times 1$

В развернутом виде для i -го наблюдения:

$$Y_i = \beta^0 + \beta^1 X^1_i + \beta^2 X^2_i + \dots + \beta^k X^k_i + \varepsilon_i$$

Ключевая интерпретация коэффициентов:

Коэффициент β^k показывает, на сколько единиц в среднем изменится зависимая переменная Y при увеличении переменной X^k на одну единицу,

при условии, что все остальные объясняющие переменные в модели остаются неизменными (*ceteris paribus*).

Пример:

$$\text{Зарплата}_i = \beta_0 + \beta_1 * \text{Образование}_i + \beta_2 * \text{Опыт}_i + \varepsilon_i$$

- β_1 : на сколько рублей изменится зарплата при увеличении образования на 1 год, **если опыт работы остается постоянным.**
- β_2 : на сколько рублей изменится зарплата при увеличении опыта на 1 год, **если уровень образования остается неизменным.**

2. Предположения классической линейной модели (МЛР)

Для множественной регрессии сохраняются все предположения КЛМР, но некоторые требуют уточнения:

1. **Линейность по параметрам:** $y = X\beta + \varepsilon$
2. **Случайность выборки:** наблюдения (X_i, Y_i) независимы и одинаково распределены
3. **Строгая экзогенность:** $E[\varepsilon_i | X] = 0$
4. **Отсутствие совершенной мультиколлинеарности:** ни один из регрессоров не является константой и не может быть точно выражен как линейная комбинация других регрессоров
5. **Гомоскедастичность:** $\text{Var}(\varepsilon_i | X) = \sigma^2$
6. **Отсутствие автокорреляции:** $\text{Cov}(\varepsilon_i, \varepsilon_j | X) = 0$ для $i \neq j$

3. Оценка параметров методом наименьших квадратов (МНК)

Как и в парном случае, мы минимизируем сумму квадратов остатков:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - b^0 - b^1 X_i^1 - \dots - b^K X_i^K)^2$$

Решение в матричной форме (формула гарантирует минимум при выполнении предположений):

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Геометрически МНК-оценка \mathbf{b} проецирует вектор \mathbf{y} на подпространство, натянутое на столбцы матрицы \mathbf{X} (Рисунок 4.1).

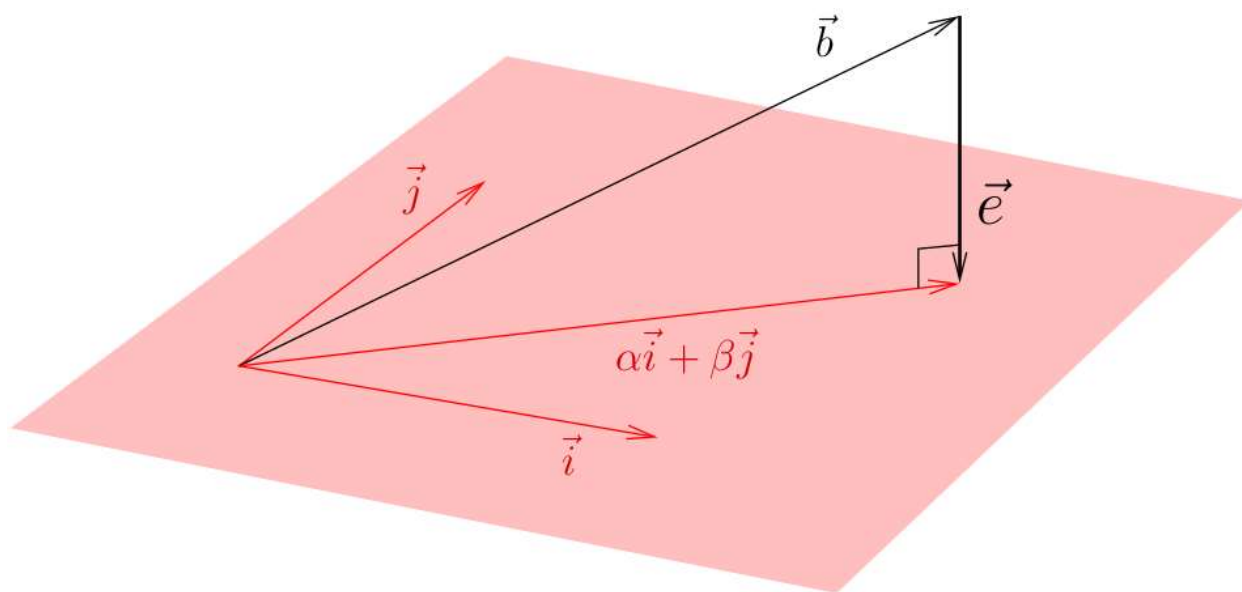


Рисунок 4.1. Геометрическая интерпретация МНК в 3D

4. Проблема мультиколлинеарности

Что это? Высокая корреляция между двумя или более объясняющими переменными.

Пример: Включение в модель и общего стажа, и стажа в текущей должности.

Чем опасна?

- Оценки коэффициентов остаются **несмещенными**, но их **дисперсии сильно возрастают**
- Коэффициенты становятся **неустойчивыми**: небольшие изменения в данных приводят к большим изменениям в оценках
- Затрудняется интерпретация: сложно выделить вклад каждой из коррелированных переменных

Диагностика:

- Высокий R^2 , но низкая t-статистика для коэффициентов
- Коэффициенты имеют неверный знак с точки зрения теории
- **Фактор инфляции дисперсии (VIF):** $VIF = 1 / (1 - R^2_j)$, где R^2_j — это R^2 регрессии j-го регрессора на все остальные. $VIF > 10$ указывает на серьезную проблему.

Как бороться?

- Убрать одну из коррелированных переменных
- Использовать главные компоненты
- Увеличить объем выборки

5. Коэффициент детерминации R^2 и скорректированный R^2

R^2 - доля дисперсии Y , объясненная моделью:

$$R^2 = ESS/TSS = 1 - (RSS/TSS)$$

Проблема: R^2 всегда увеличивается при добавлении нового регрессора, даже если он статистически незначим.

Решение: скорректированный R^2 (\bar{R}^2) — штрафует за добавление неинформативных переменных:

$$\begin{aligned} R^2 &= 1 - [(RSS/(n - k - 1))/(TSS/(n - 1))] \\ &= 1 - (1 - R^2) * (n - 1)/(n - k - 1) \end{aligned}$$

где:

- n — число наблюдений
- k — число регрессоров (без константы)

\bar{R}^2 используется для сравнения моделей с разным числом регрессоров.

6. Пример: Детерминанты заработной платы

Оценим модель:

$$\text{Зарплата} = \beta_0 + \beta_1 * \text{Образование} + \beta_2 * \text{Опыт} + \beta_3 * \text{Пол} + \varepsilon$$

где Пол — бинарная переменная (1-мужчина, 0-женщина).

Результаты оценки:

$$\text{Зарплата} = -20 + 4.5 * \text{Образование} + 1.2 * \text{Опыт} + 8.0 * \text{Пол}$$

Интерпретация:

- $\beta_1 = 4.5$: при увеличении образования на 1 год зарплата растет на 4.5 тыс. руб., *при неизменных опыте и поле*
- $\beta_2 = 1.2$: при увеличении опыта на 1 год зарплата растет на 1.2 тыс. руб., *при неизменных образовании и поле*
- $\beta_3 = 8.0$: Мужчины в среднем получают на 8.0 тыс. руб. больше женщин, *с одинаковым образованием и опытом*
-

Резюме

1. **Множественная регрессия** позволяет оценить "чистый" эффект каждого фактора
2. **Интерпретация коэффициентов** всегда включает условие *ceteris paribus*
3. **Мультиколлинеарность** — серьезная проблема, ухудшающая качество оценок
4. **Скорректированный R^2** — более адекватная мера качества модели, чем обычный R^2
5. **Матричная форма записи** компактна и удобна для вывода формул

На следующей лекции: Мы научимся проводить **статистические выводы** в множественной регрессии: проверять гипотезы о коэффициентах и значимость модели в целом.

Вопросы для самопроверки:

1. Почему оценка влияния образования на зарплату в парной регрессии может быть смещенной?
2. В модели для спроса на кофе коэффициенты при цене на кофе и цене на чай оказались незначимы. О чем это может говорить?
3. При добавлении в регрессию новой переменной R^2 вырос, а \bar{R}^2 уменьшился. Какой вывод следует сделать?