# Лекция 3

# Линейная регрессия

## 1. Введение в линейную регрессию

Линейная регрессия — это один из наиболее базовых и мощных инструментов в арсенале машинного обучения и статистики, используемый для выявления и моделирования линейных зависимостей между переменными. Основная цель линейной регрессии — предсказать значение зависимой переменной, также называемой целевой переменной, основываясь на значениях одной или нескольких независимых переменных, или предикторов. Линейная регрессия предполагает, что изменения в независимых переменных пропорционально влияют на изменения в зависимой переменной, что позволяет представить связь между переменными в виде линейной функции.

Модель линейной регрессии широко используется в самых различных областях. В экономике она помогает прогнозировать макроэкономические показатели, такие как инфляция и ВВП, а также анализировать связь между спросом и предложением. В социологии линейная регрессия применяется для выявления корреляций между социальными факторами, такими как образование и уровень дохода. В медицине метод используется для анализа и прогнозирования различных клинических исходов, таких как реакция пациента на лечение или вероятность развития заболевания. Простота и интерпретируемость линейной регрессии делают её универсальным инструментом для анализа данных и предсказания на основе числовых значений.

## 2. Основные понятия и математическая постановка задачи

Линейная регрессия основывается на простой математической модели, описывающей зависимость между переменными. Общий вид линейной регрессии можно записать следующим образом:

y=β0+β1x1+β2x2+...+βnxn+ϵy = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n + \epsilony=β0+β1x1+β2x2+...+βnxn+ϵ

где:

- yyy — зависимая переменная (целевая переменная),
- xix_ixi — независимые переменные,

- β0\beta_0β0 — свободный член (intercept),
- βi\beta_iβi — коэффициенты регрессии для независимых переменных,
- ϵ\epsilonϵ — случайная ошибка или остаток.

Цель метода линейной регрессии — оценить параметры β0,β1,...,βn\beta_0, \beta_1, ..., \beta_nβ0,β1,...,βn таким образом, чтобы минимизировать разницу между предсказанными значениями и реальными значениями целевой переменной.

### 3. Метод наименьших квадратов

Один из самых распространенных способов нахождения параметров βi\beta_iβi — это метод наименьших квадратов. Метод наименьших квадратов минимизирует сумму квадратов отклонений предсказанных значений от фактических:

Minimize ∑i=1m(yi−yi^)2=∑i=1m(yi−(β0+β1xi1+...+βnxin))2\text{Minimize } \sum_{i=1}^{m} (y_i - \hat{y_i})^2 = \sum_{i=1}^{m} (y_i - (\beta_0 + \beta_1x_{i1} + ... + \beta_nx_{in}))^2Minimize i=1∑m(yi−yi^)2=i=1∑m(yi−(β0+β1xi1+...+βnxin))2

где mmm — число наблюдений.

Этот метод позволяет получить коэффициенты, при которых модель будет предсказывать значения, минимально отклоняющиеся от фактических.

### 4. Множественная линейная регрессия

Множественная линейная регрессия расширяет простой линейный подход, включая несколько независимых переменных. Модель множественной линейной регрессии выглядит следующим образом:

y=β0+β1x1+β2x2+...+βnxn+ϵy = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n + \epsilony=β0+β1x1+β2x2+...+βnxn+ϵ

Множественная линейная регрессия позволяет моделировать более сложные взаимосвязи и может быть полезна, когда зависимая переменная зависит от нескольких факторов. Однако наличие большого числа переменных может приводить к повышению сложности модели и потенциальному переобучению, особенно в условиях малого набора данных.

### 5. Основные предположения линейной регрессии

Для успешного применения линейной регрессии требуется соблюдение ряда предположений:

- **Линейность**: Предполагается, что зависимость между независимыми переменными и целевой переменной линейна.
- **Независимость ошибок**: Ошибки должны быть независимы друг от друга.
- **Гомоскедастичность**: Ошибки имеют постоянную дисперсию при любых значениях независимых переменных.
- **Нормальность ошибок**: Ошибки следуют нормальному распределению.

Если эти предположения не выполняются, качество модели может снизиться, и могут потребоваться альтернативные методы.

### 6. Методы оценки качества модели

Оценка качества линейной регрессии может проводиться с использованием различных метрик. Наиболее распространённые метрики включают:

- **Среднеквадратическая ошибка (MSE)**: Это среднее значение квадратов отклонений между предсказанными и реальными значениями:

MSE=1m∑i=1m(yi−yi^)2MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y_i})^2MSE=m1i=1∑m(yi−yi^)2

- **Средняя абсолютная ошибка (MAE)**: Среднее значение абсолютных отклонений предсказаний от фактических значений:

MAE=1m∑i=1m|yi−yi^|MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y_i}|MAE=m1i=1∑m|yi−yi^|

- **Коэффициент детерминации R2R^2R2**: Определяет долю дисперсии, объясненную моделью. Значение R2R^2R2 приближается к 1 для хорошо объясняющей модели и к 0 для модели, не объясняющей дисперсию зависимой переменной.

### 7. Регуляризация: Ridge и Lasso регрессия

Регуляризация используется для снижения вероятности переобучения модели и улучшения её устойчивости. Два основных подхода к регуляризации в линейной регрессии — это Ridge (гребневая регрессия) и Lasso (Least Absolute Shrinkage and Selection Operator).

- **Ridge регрессия** добавляет штраф за большие значения коэффициентов, минимизируя следующую функцию потерь:

Minimize ∑i=1m(yi−yi^)2+λ∑j=1nβj2\text{Minimize } \sum_{i=1}^{m} (y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{n} \beta_j^2Minimize i=1∑m(yi−yi^)2+λj=1∑nβj2

- **Lasso регрессия** добавляет штраф на сумму абсолютных значений коэффициентов, минимизируя:

Minimize ∑i=1m(yi−yi^)2+λ∑j=1n|βj|\text{Minimize } \sum_{i=1}^{m} (y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{n} |\beta_j|Minimize i=1∑m(yi−yi^)2+λj=1∑n|βj|

**8. Применение линейной регрессии**

Линейная регрессия — универсальный метод, нашедший применение в самых разных областях науки и практики. Её способность моделировать и объяснять линейные зависимости между переменными делает её незаменимым инструментом для анализа и прогнозирования. Рассмотрим несколько примеров применения линейной регрессии в различных областях:

- **Экономика**: Линейная регрессия широко используется для анализа и прогнозирования экономических показателей. Например, она позволяет моделировать связь между ценой на товар и его спросом. В микроэкономике модель может использоваться для понимания влияния различных факторов, таких как цена и доход потребителей, на потребление товаров. В макроэкономике линейная регрессия применяется для прогнозирования темпов роста ВВП, инфляции, уровня занятости и других показателей. Экономисты также используют линейную регрессию для оценки эффектов государственных политик и анализа финансовых рынков.
- **Медицина**: В медицинской статистике и биостатистике линейная регрессия применяется для анализа факторов, влияющих на риск развития различных заболеваний. Например, она позволяет изучать влияние возраста, уровня холестерина, давления и других факторов на вероятность сердечно-сосудистых заболеваний. Линейные модели также применяются для оценки эффективности лечения, где в зависимости от факторов, таких как дозировка лекарства или продолжительность терапии, можно предсказать вероятный результат лечения или побочные эффекты. В эпидемиологии

линейная регрессия помогает моделировать распространение заболеваний и определять факторы риска для здоровья населения.

- **Маркетинг**: В маркетинговых исследованиях линейная регрессия используется для оценки эффективности рекламных кампаний и выявления факторов, влияющих на продажи. Например, компании могут использовать линейные модели для анализа зависимости между затратами на маркетинг и объемом продаж, что позволяет оптимизировать бюджет рекламной кампании. Линейная регрессия также может быть применена для анализа покупательских предпочтений и поведения, что помогает компаниям разрабатывать более целевые стратегии для увеличения конверсии и удержания клиентов.
- **Физика**: В физике линейная регрессия помогает анализировать экспериментальные данные и выявлять линейные зависимости между физическими величинами. Например, закон Гука, описывающий связь между силой упругости и деформацией пружины, является примером линейной зависимости, которую можно проанализировать с помощью линейной регрессии. Физики часто используют линейные модели для обработки результатов экспериментов, чтобы оценить неизвестные параметры и проверить гипотезы о физических явлениях.

Эти примеры иллюстрируют, насколько разнообразны области применения линейной регрессии. Её способность выявлять и объяснять взаимосвязи между переменными делает её важным инструментом для анализа данных и принятия решений, основанных на количественных оценках.

### 9. Ограничения линейной регрессии

Несмотря на широкое применение и универсальность, линейная регрессия имеет ряд ограничений, которые могут значительно повлиять на точность и интерпретацию модели. Понимание этих ограничений важно для того, чтобы адекватно оценивать результаты модели и выбирать подходящие методы в зависимости от конкретной задачи. Основные ограничения линейной регрессии включают следующие аспекты:

- **Линейность предположения**: Основное предположение линейной регрессии заключается в том, что зависимость между независимыми переменными и целевой переменной является линейной. Это значит, что модель линейной регрессии плохо справляется с задачами, где отношения между переменными сложные и нелинейные. В реальных данных линейная зависимость встречается редко, особенно в задачах с большим количеством факторов и сложными взаимодействиями. Если зависимость нелинейна, то линейная регрессия может давать смещенные оценки и существенно занижать точность модели. В таких случаях целесообразно использовать более сложные методы, например, полиномиальную регрессию, метод опорных векторов или нейронные сети.
- **Чувствительность к выбросам**: Линейная регрессия весьма чувствительна к выбросам, или аномальным значениям, в данных. Наличие выбросов может значительно искажать оценку параметров модели, так как линейная регрессия минимизирует сумму квадратов отклонений. Это приводит к тому, что выбросы, находящиеся далеко от общего тренда данных, оказывают непропорционально сильное влияние на модель, смещая её линию. Чтобы уменьшить влияние выбросов, можно использовать методы очистки данных или использовать более устойчивые методы регрессии, такие как регрессия на основе медианы (например, метод наименьших абсолютных отклонений) или регрессия с весами для выбросов.
- **Корреляция между переменными (мультиколлинеарность)**: Мультиколлинеарность — это ситуация, когда независимые переменные сильно коррелируют между собой. Это приводит к нестабильности оценки коэффициентов линейной регрессии, поскольку модель затрудняется определить вклад каждой переменной в предсказание. В случае высокой мультиколлинеарности, небольшие изменения в данных могут привести к значительным колебаниям в оценке коэффициентов, что усложняет интерпретацию модели. Для решения этой проблемы можно использовать методы устранения мультиколлинеарности, такие как исключение высококоррелированных признаков или применение регуляризованных моделей, например, Lasso или Ridge регрессии, которые ограничивают или штрафуют коэффициенты, уменьшая влияние мультиколлинеарности.

Эти ограничения указывают на то, что линейная регрессия не всегда является оптимальным выбором, особенно при работе с большими, сложными и разнородными данными. В таких случаях анализ данных и применение методов, устойчивых к данным ограничениям, помогает повысить качество модели и точность предсказаний.

## 10. Заключение

Линейная регрессия является мощным инструментом для анализа данных и построения прогнозов. Она проста в реализации и интерпретации, но требует внимательного подхода к проверке предположений и оценке качества модели. В сочетании с методами регуляризации и правильной предобработкой данных линейная регрессия способна решать широкий круг задач в бизнесе, науке и технике.