


# Регрессионный анализ (подготовка данных, расчет, интерпретация результатов)



- Это группа методов, направленных на выявление и математическое выражение тех изменений и зависимостей, которые имеют место в системе случайных величин.
- Если такая система моделирует педагогическую, психологическую, биологическую и т. п., то, следовательно, путем регрессионного анализа выявляются и математически выражаются явления научного эксперимента и зависимости между ними
- Регрессионный анализ (регрессивное исследование) в психологии - это метод математической статистики, который позволяет изучать зависимость среднего значения любой величины от вариаций другой величины или нескольких величин

# Основные цели регрессионного анализа

- **Предсказание значения зависимой переменной** с помощью независимых переменных.
- . Определение вклада отдельных независимых переменных в вариацию зависимой переменной

# задачи регрессионного анализа

- Первая задача – выявить факт изменчивости изучаемого явления при определенных, но не всегда четко фиксированных условиях.
- Вторая задача – выявить тенденцию как периодическое изменение признака.
- Третья задача – это выявление закономерности, выраженной в виде корреляционного уравнения (регрессии).

# Линейная модель парной регрессии

- Регрессия – функция, позволяющая по величине одного коррелируемого признака определить среднюю величину другого признака.

# Основные этапы регрессионного анализа

Первый этап. Предположение. На этом этапе происходит выбор формы связи между переменными (модель).

Второй этап. Параметризация – происходит оценка значений параметра в выбранной формуле статистической связи. Форма связи (функция) линейная, нелинейная.

Третий этап. Проверка надёжности полученных оценок. На этом этапе осуществляются следующие тесты: F-тест (проверка статистической значимости выбранной формы связи), t-тест (проверка статистической значимости найденных числовых значений параметра).

- В результате анализа статистических данных, выбора и построения модели последовательно выполняются все три этапа.
- Замечание. На одну переменную, входящую в модель, должно приходиться не менее 6–7 объектов из рассматриваемой выборки.

# Предположения линейной модели

Линейный регрессионный анализ основан на шести фундаментальных предположениях:

1. Переменные показывают линейную зависимость;
2. Независимая переменная не случайна;
3. Значение невязки (ошибки) равно нулю;
4. Значение невязки постоянно для всех наблюдений;
5. Значение невязки не коррелирует по всем наблюдениям;
6. Остаточные значения подчиняются нормальному распределению.

# Построение простой линейной регрессии

- Простая линейная модель выражается с помощью следующего уравнения:

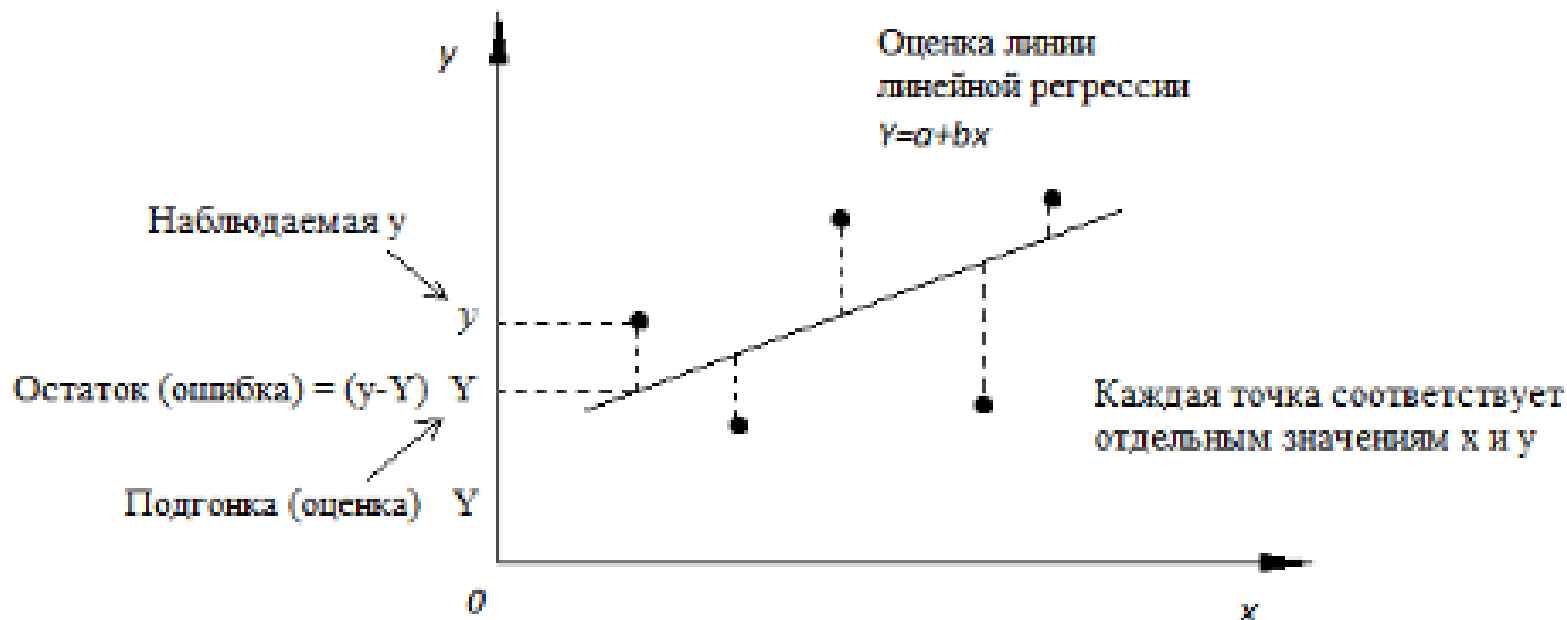
$$Y = a + bX$$

Где:

- $Y$  — зависимая переменная
- $X$  — независимая переменная (объясняющая)
- $a$  — свободный член (сдвиг по оси  $OY$ )
- $b$  — угловой коэффициент. Он указывает на поведение кривой (убывает или возрастает, угол между с осью)

$a$  и  $b$  называют коэффициентами линейной регрессии. В их нахождении и заключается основная задача.





Коэффициенты  $a$  и  $b$  определяются с помощью метода наименьших квадратов, который минимизирует сумму квадратов разностей между наблюдаемыми значениями зависимой переменной и предсказанными значениями, полученными с помощью уравнения регрессии.

- Простая линейная регрессия также позволяет оценить статистическую значимость связи между переменными с помощью t-теста и рассчитать коэффициент детерминации (R-квадрат), который показывает, насколько хорошо модель объясняет вариацию в зависимой переменной.
- Простая линейная регрессия может быть полезна для прогнозирования значений зависимой переменной на основе независимой переменной, а также для понимания и изучения взаимосвязей между переменными.

# основные задачи регрессионного анализа

- **построение уравнения регрессии**, т.е. нахождение вида зависимости между результатным показателем и независимыми факторами  $x_1, x_2, \dots, x_n$ .
- **оценка значимости полученного уравнения**, т.е. определение того, насколько выбранные факторные признаки объясняют вариацию признака  $y$

# Основные определения

- **Зависимая переменная (Y)** — это переменная, описывающая процесс, который мы пытаемся предсказать или понять.
- **Независимые переменные (X)** это переменные, используемые для моделирования или прогнозирования значений зависимых переменных.
- Зависимая переменная - это функция независимых переменных.
- **Коэффициенты регрессии ( $\beta$ )** — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.
- **Невязки.** Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные

# Регрессионный анализ (достоинства и недостатки)

Регрессионный анализ нельзя использовать для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для применения этого вида анализа

## Достоинства

- 1 Простота вычислительных алгоритмов.
- 2 Наглядность и интерпретируемость результатов (для линейной модели).

## Недостатки

- **1 Невысокая точность** прогноза (в основном - интерполяция данных).
- **2 Субъективный характер выбора вида конкретной зависимости** (формальная подгонка модели под эмпирический материал).
- 3 Отсутствие объяснительной функции (**невозможность объяснения причинно-следственной связи**).

	корреляционный анализ	регрессионный анализ
Задачи	исследование тенденций взаимного изменения двух или более случайных величин	построение зависимости математического ожидания одной или нескольких случайных величин от одной или нескольких неслучайных величин
Использование	Связь как синхронность (согласованность)	Связь как зависимость (влияние), (причинно-следственные связи)
Коэффициент	-Принимает значения в диапазоне от -1 до +1; -Безразмерная величина; -Показывает силу связи между признаками; -Знак коэффициента говорит о направлении связи.	-Может принимать любые значения; -Привязан к единицам измерения обоих признаков; -Показывает структуру связи между признаками; -Знак коэффициента говорит о направлении связи.

# Множественная линейная регрессия

Множественная линейная регрессия – это статистический метод, который позволяет исследовать взаимосвязь между одной зависимой переменной и двумя или более независимыми переменными.

Он является расширением простой линейной регрессии, где есть только одна независимая переменная.

- В множественной линейной регрессии модель представляется уравнением:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n + \varepsilon$$

- где  $Y$  – зависимая переменная,
- $X_1, X_2, \dots, X_n$  – независимые переменные,
- $b_0, b_1, b_2, \dots, b_n$  – коэффициенты регрессии,
- $\varepsilon$  – ошибка.

# Цель множественной линейной регрессии

- – оценить значения коэффициентов регрессии ( $b_0, b_1, b_2, \dots, b_n$ ), чтобы понять, как каждая независимая переменная влияет на зависимую переменную.
- Коэффициенты регрессии показывают, насколько изменится зависимая переменная при изменении соответствующей независимой переменной, при условии, что все остальные переменные остаются постоянными.



- Коэффициент детерминации (R-квадрат), который показывает, насколько хорошо модель объясняет вариацию в зависимой переменной.
- Множественная линейная регрессия может быть полезна для прогнозирования значений зависимой переменной на основе независимых переменных, а также для понимания и изучения взаимосвязей между переменными.

- Оценка параметров регрессии позволяет определить, насколько каждая независимая переменная влияет на зависимую переменную. Значения коэффициентов регрессии показывают, насколько изменится зависимая переменная при изменении на единицу каждой независимой переменной, при условии, что все остальные переменные остаются постоянными.
- Оценка параметров регрессии также позволяет оценить статистическую значимость каждого коэффициента. Для этого используется  $t$ -статистика и соответствующий уровень значимости. Если  $p$ -значение меньше выбранного уровня значимости (обычно 0.05), то можно считать, что коэффициент статистически значимо отличается от нуля.

# Коэффициент детерминации

- Коэффициент детерминации (R-квадрат) показывает, какой процент вариации зависимой переменной объясняется независимыми переменными в модели. Он принимает значения от 0 до 1, где 0 означает, что модель не объясняет вариацию, а 1 означает, что модель полностью объясняет вариацию.
- Чем ближе коэффициент детерминации к 1, тем лучше модель описывает зависимость между переменными. Однако, следует помнить, что высокий коэффициент детерминации не всегда означает, что модель является хорошей или предсказательной.
- Проверка значимости модели является важным шагом в регрессионном анализе, так как позволяет определить, насколько хорошо модель описывает данные и является ли она статистически значимой.

# Диагностика модели

- Диагностика модели в регрессионном анализе включает в себя проверку предпосылок и оценку качества модели. Это важный шаг, который позволяет убедиться в адекватности и надежности полученных результатов.

# Проверка предпосылок

- **Линейность:** Проверка, является ли отношение между зависимой переменной и независимыми переменными линейным. Для этого можно построить графики рассеяния и проверить их линейность.
- **Нормальность:** Проверка, являются ли остатки модели нормально распределенными. Для этого можно построить гистограмму остатков и провести тест на нормальность, например, тест Шапиро-Уилка.
- **Гомоскедастичность:** Проверка, являются ли остатки модели однородными по дисперсии. Для этого можно построить график остатков от предсказанных значений и провести тест на гомоскедастичность, например, тест Бройша-Пагана.
- **Отсутствие мультиколлинеарности:** Проверка, нет ли сильной корреляции между независимыми переменными. Для этого можно построить матрицу корреляции и провести тест на мультиколлинеарность, например, тест Варнера.

# Оценка качества модели

- **Оценка коэффициентов:** Оценка значимости и величины коэффициентов регрессии. Для этого используются статистические тесты, такие как t-тест или F-тест.
- **Оценка качества подгонки:** Оценка того, насколько хорошо модель подгоняется под данные. Для этого используются различные метрики, такие как коэффициент детерминации ( $R^2$ ), среднеквадратическая ошибка (MSE) и другие.
- **Проверка статистической значимости модели:** Проверка, является ли модель статистически значимой в целом. Для этого используется F-тест.
- Важно отметить, что диагностика модели является итеративным процессом, и в случае выявления нарушений предпосылок или низкого качества модели, необходимо провести коррекцию или выбрать другую модель.

# Интерпретация результатов

## Коэффициенты регрессии

- Коэффициенты регрессии представляют собой числовые значения, которые показывают, как изменение одной переменной влияет на изменение зависимой переменной.
- Например, в случае простой линейной регрессии, коэффициент наклона (бета-коэффициент) показывает, насколько изменится зависимая переменная при изменении независимой переменной на единицу.
- Интерпретация коэффициентов регрессии должна быть основана на контексте и предметной области исследования.

- **Значимость коэффициентов**
- При интерпретации результатов регрессионного анализа важно учитывать статистическую значимость коэффициентов. Если коэффициент является статистически значимым, это означает, что существует значимая связь между независимой и зависимой переменными. Если коэффициент не является статистически значимым, это может указывать на отсутствие связи или недостаточную выборку для обнаружения связи.
- **Коэффициент детерминации**
- Коэффициент детерминации ( $R^2$ -квадрат) показывает, насколько хорошо модель объясняет изменчивость зависимой переменной. Интерпретация коэффициента детерминации может быть следующей: если  $R^2$ -квадрат равен 0.8, это означает, что 80% изменчивости зависимой переменной объясняется независимыми переменными в модели, а оставшиеся 20% могут быть объяснены другими факторами, которые не были включены в модель.



- **Предсказание и прогнозирование**
- Регрессионный анализ также может использоваться для предсказания значений зависимой переменной на основе значений независимых переменных. Интерпретация результатов предсказания должна быть основана на доверительных интервалах и стандартных ошибках предсказания. Прогнозирование, с другой стороны, предполагает использование модели для предсказания значений зависимой переменной вне диапазона наблюдаемых данных.
- В целом, интерпретация результатов регрессионного анализа требует внимательного анализа и понимания статистических показателей, а также учета контекста и предметной области исследования. Это поможет сделать осмысленные выводы и принять информированные решения на основе полученных результатов.