Регрессионный анализ

$$Y=a+b*X$$

Регрессионный анализ

- Впервые термин употреблен в работе Pearson (1908)
- Анализ связи между несколькими независимыми переменными (регрессорами или предикторами) и зависимой переменной

Цели регрессионного анализа

- Определение наличия и характера (математического уравнения, описывающего зависимость) связи между переменными
- Определение степени детерминированности вариации критеральной переменной предикторами
- Предсказать значение зависимой переменной с помощью независимой
- Определить вклад независимых переменных в вариацию зависимой

Условия применения

- Использование метрических переменных
- Равенство условных дисперсий: $D(Y / X) = const$;
- Независимость ошибок от предикторов и нормального их распределения с нулевым средним и постоянной дисперсией;
- Попарное нормальное распределение всех признаков модели;
- Независимость предикторов между собой
- Достаточное количество наблюдений (обычно >15, в зависимости от конкретного характера распределений наблюдений и сложности искомой зависимости)

Уравнение регрессии

Y=a+b*X; где:

- Y – зависимая переменная,
- a - константа
- b - угловой коэффициент
- X – независимая переменная

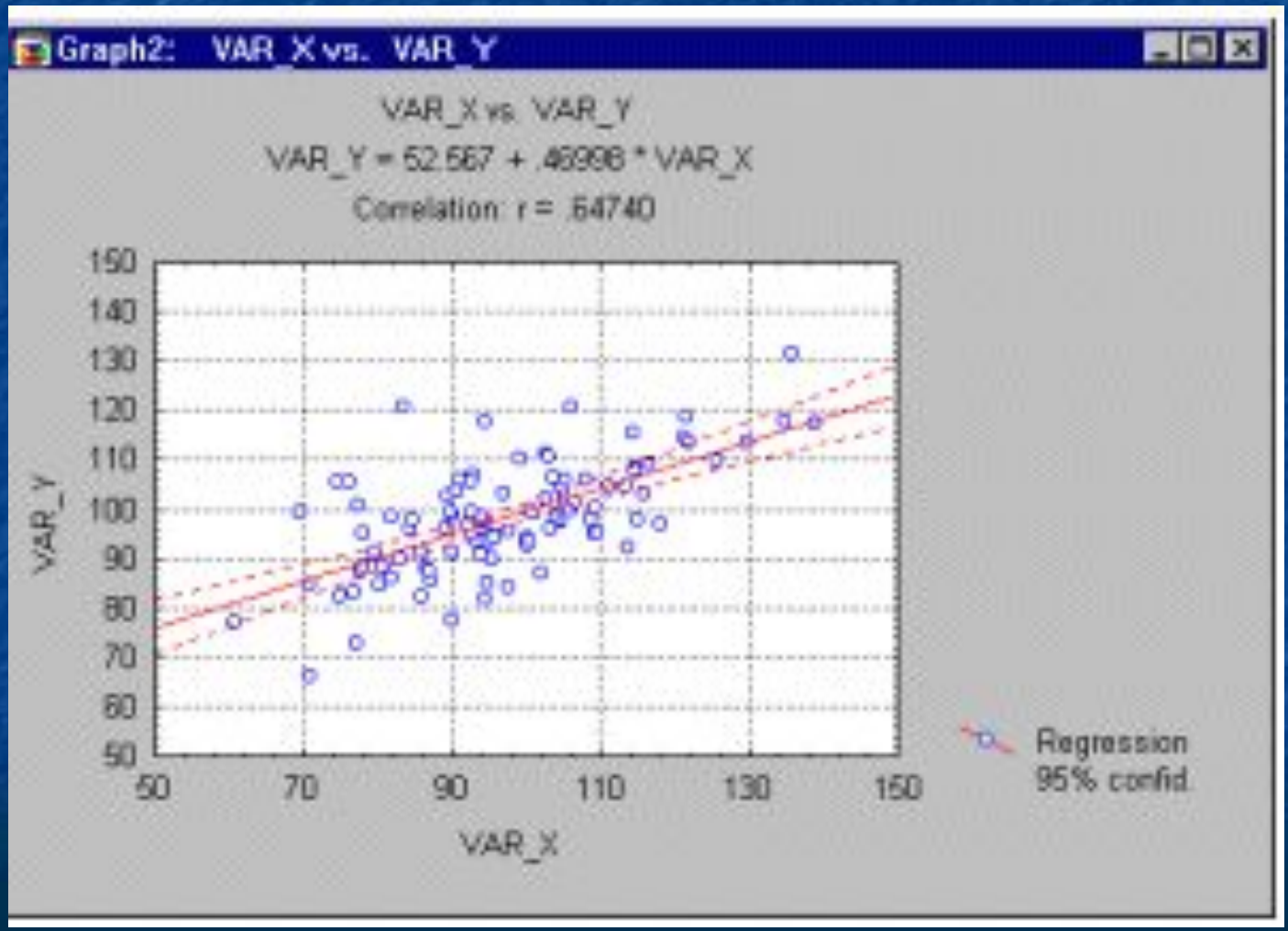
Для многомерной регрессии:

Y = a + b1*X1 + b2*X2 + … + bp*Xp

Метод наименьших квадратов

- Цель - минимизировать квадраты отклонений линии регрессии от наблюдаемых точек.
- По этим данным строим диаграмму рассеяния

Диаграмма рассеяния



Регрессионные коэффициенты (B-коэффициенты)

- Это независимые вклады каждой независимой переменной в предсказание зависимой переменной:

  переменная X1 коррелирует с переменной Y после учета влияния всех других независимых переменных (частная корреляция)

Пример

- Успеваемость = 1+.02*IQ, где:
- a = 1
- b = 0,02
- IQ – независимая переменная

- При IQ=130:
- Успеваемость = 1+.02*130 = 3,6

Остаток

- Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется остатком.
- Чем меньше разброс значений (дисперсия) остатков около линии регрессии по отношению к общему разбросу значений, тем лучше прогноз

Остаточная дисперсия и коэффициент детерминации R-квадрат

- Если связь между переменными X и Y отсутствует, то отношение остаточной изменчивости переменной Y к исходной дисперсии равно 1.0.
- Если X и Y жестко связаны, то остаточная изменчивость отсутствует, и отношение дисперсий будет равно 0.0.
- В большинстве случаев отношение будет лежать между экстремальными значениями, т.е. между 0.0 и 1.0.
- 1.0 минус это отношение называется R-квадратом или коэффициентом детерминации

Коэффициент множественной корреляции R

- Это неотрицательная величина, принимающая значения между 0 и 1.
- Если B-коэффициент положителен, то связь этой переменной с зависимой переменной положительна
- Если B-коэффициент отрицателен, то и связь носит отрицательный характер.
- Конечно, если B-коэффициент равен 0, связь между переменными отсутствует.

Спасибо за внимание