Регрессия (Regression)

В теории вероятности и математической статистике это зависимость математического ожидания случайной величины от одной или нескольких других случайных величин.

В отличие от чисто функциональной зависимости $y = f(x)$, где каждому значению независимой переменной x соответствует единственное значение зависимой переменной y, регрессионная зависимость предполагает, что каждому значению переменной x могут соответствовать различные значения y, обусловленные случайной природой зависимости.

Если некоторому значению величины x_i соответствует набор значений величин $y_{i1}, y_{i2}, \ldots, y_{in}$, то зависимость средних арифметических:

$\bar{y}_i = \frac{y_{i1}+y_{i2}+\cdots+y_{in}}{n_i}$

от x_i и является регрессией в статистическом понимании данного термина.

Типичным примером регрессионной зависимости может быть зависимость между ростом и весом человека. В большинстве случае вес пропорционален росту, но фактически большой рост не всегда означает большой вес. Иными словами, у роста, например, 175 см. может наблюдаться несколько значений веса, скажем 69, 78 и 86 кг. Тогда зависимость между ростом и средним значением указанных весов будет являться регрессионной.

Изучение регрессии в теории вероятностей основано на том, что случайные величины X и Y, имеющие совместное распределение вероятностей, связаны статистической зависимостью: при каждом фиксированном значении $X = x$, величина Y является случайной величиной с определенным (зависящим от значения x) условным распределением вероятностей.

Регрессия величины Y по величине X определяется условным математическим ожиданием Y, вычисленным при условии, что $X = x : E(Y|x) = u(x)$.

Уравнение $y = u(x)$ называется **уравнением регрессии**, а соответствующий график — линией регрессии Y по X. Точность, с которой уравнение Y по X отражает изменение Y в среднем при изменении x, измеряется условной дисперсией D величины Y, вычисленной для каждого значения $X = x : D(Y|x) = D(x)$.

Если $D(x) = 0$ при всех значениях x, то можно достоверно утверждать, что Y и X связаны строгой функциональной зависимостью $Y = u(X)$. Если $D(x) = 0$ при всех значениях x и $u(x)$ не зависит от x, то говорят, что регрессионная зависимость Y по X отсутствует.

Линии регрессии обладают следующим **замечательным свойством**: среди всех действительных функций $f(X)$ минимум математического ожидания $E[Y— f(X)]^2$ достигается для функции $f(x) = u(X)$.

Это означает, что регрессия Y по X дает наилучшее в указанном смысле представление величины Y по величине X. Это свойство позволяет использовать регрессию для предсказания величины Y по X.

Иными словами, если значение Y непосредственно не наблюдается и эксперимент позволяет регистрировать только X, то в качестве прогнозируемого значения Y можно использовать величину $Y = u(X)$.

Наиболее простым является случай, когда регрессионная зависимость Y по X линейна, т.е. $E(Y|x) = b_0 + b_1 x$, где b_0 и b_1 — коэффициенты регрессии. На практике обычно коэффициенты регрессии в уравнении $y = u(x)$ неизвестны, и их оценивают по наблюдаемым данным.

Регрессия широко используется в аналитических технологиях при решении различных бизнес-задач, таких как прогнозирование (продаж, курсов валют и акций), оценивание различных бизнес-показателей по наблюдаемым значениям других показателей (скоринг), выявление зависимостей между показателями и т.д.

В Loginom существует специализированный обработчик логистическая регрессия, с помощью которого можно оценивать вероятность того, что событие наступит для конкретного объекта испытания (больной/здоровый, возврат кредита/дефолт и т.д.). И обработчик линейная регрессия, который может использоваться для решения различных задач, например, прогнозирования и численного предсказания.

Логистическая регрессия является традиционным статистическим инструментом для расчета коэффициентов (баллов) скоринговой карты на основе накопленной кредитной истории. Подробнее в статье «Логистическая регрессия и ROC-анализ — математический аппарат».

О прикладном применении логистической регрессии в двух областях — диагностика заболеваний и оценка кредитоспособности физических лиц узнайте в статье «Применение логистической регрессии в медицине и скоринге».