АНАЛИЗ РЕГРЕССИОННЫЙ - группа методов статистического анализа данных, предназначенных для исследования причинных связей между количественными переменными (см.). В общем виде, регрессионную зависимость можно представить в виде функции $y = f(x_1, x_2, \dots, x_k)$, где y - переменная зависимая (см.), $x_1, x_2, \dots, x_k$ - переменные независимые (см.) или предикторы. Если функция f является линейной, говорят о линейной регрессии, если нет - о регрессии нелинейной (см.). Наиболее популярными нелинейными моделями в социальных науках являются логарифмическая $y = b \ln(x) + a$ и экспоненциальная $y = e^{bx+a}$.

В самом простом случае уравнение линейной регрессии включает две переменные $y = bx+a$. Если в уравнении используется несколько независимых переменных, регрессия называется множественной. Уравнение множественной линейной регрессии имеет вид $y = \sum b_i x_i + a$.

В настоящее время используются два подхода к интерпретации коэффициентов линейной регрессии b_i . Согласно первому из них, b_i представляет собой величину, на которую изменится предсказанное по модели значение $\hat{y} = \sum b_i x_i + a$ при увеличении значения независимой переменной x_i на одну единицу измерения; согласно второму - величину, на которую, в среднем, изменяется значение переменной y при увеличении независимой переменной x_i на единицу. Свободный член уравнения регрессии a равен предсказанному значению зависимой переменной $\hat{y}$ в случае, когда все независимые переменные $x_i = 0$.

Теснота регрессионной связи измеряется коэффициентом детерминации (см.), который интерпретируется как доля дисперсии зависимой переменной y , объясненная независимыми переменными $x_1, x_2, \dots, x_k$ (см.).

Дисперсия объясненная). Для парной линейной регрессии коэффициент детерминации равен квадрату коэффициента линейной корреляции Пирсона (см.) r^2 , для множественной линейной регрессии - квадрату коэффициента корреляции множественной (см.) R^2 .

Современные регрессионные техники позволяют включать в уравнение не только количественные, но и качественные, в первую очередь дихотомические, переменные. Независимые дихотомические переменные включаются в регрессионные модели на общих основаниях. Номинальные переменные с числом градаций $k > 2$ подвергаются процедуре дитту-кодирования (см.), после чего могут быть включены в модель в виде набора $(k - 1)$ независимых дихотомических переменных.

Если дихотомическая переменная ($y = 1$ - "успех", $y = 0$ - "неуспех") является зависимой,

то уравнение $p = \sum b_i x_i + a$ - наиболее простой способ предсказать вероятность "успеха" (того, что событие произойдет или объект обладает требуемым качеством). К сожалению, эта простая модель отличается неустойчивостью и низкой достоверностью результатов.

Несколько более сложной, но и перспективной, является модель логистической (логит-) регрессии, в которой вероятность успеха предсказывается по формуле . Эту модель можно представить более удобной для понимания формулой $\ln(p/1 - p) = \sum b_i x_i + a$. Она представляет логарифм отношения вероятности "успеха" p к вероятности "неуспеха" $1 - p$ как линейную функцию независимых переменных $x_1, x_2, \dots, x_k$.

Еще одна популярная модель для дихотомической зависимой переменной - пробит-регрессия. В ней вероятность успеха предсказывается по формуле , где z - случайная величина, распределенная по нормальному закону Гаусса.

Следующим этапом развития статистического моделирования явилось обобщение логит- и пробит-моделей на порядковые и номинальные зависимые переменные с числом градаций больше двух. Они называются мультиномиальными и порядковыми логит- и пробит-моделями и пользуются большой популярностью благодаря своей высокой эффективности и устойчивости, а также развитию программных средств статистического анализа данных.

О.В. Терещенко