Раздел 3. Методы анализа данных в маркетинге.

Тема 3.1. Регрессионный анализ (2 часа)

Регрессионный анализ (regression analysis) – это метод изучения статистической взаимосвязи между одной зависимой количественной зависимой переменной от одной или нескольких независимых количественных переменных. Зависимая переменная в регрессионном анализе называется результирующей, а переменные факторы – предикторами или объясняющими переменными.

Взаимосвязь между средним значением результирующей переменной и средними значениями предикторов выражается в виде уравнения регрессии. Уравнение регрессии – математическая функция, которая подбирается на основе исходных статистических данных зависимой и объясняющих переменных. Чаще всего используется линейная функция. В этом случае говорят о линейном регрессионном анализе.

Регрессионный анализ очень тесно связан с корреляционным анализом. В корреляционном анализе исследуется направление и теснота связи между количественными переменными. В регрессионном анализе исследуется форма зависимости между количественными переменными. Т.е. фактически оба метода изучают одну и ту же взаимосвязь, но с разных сторон, и дополняют друг друга. На практике корреляционный анализ выполняется перед регрессионным анализом. После доказательства наличия взаимосвязи методом корреляционного анализа можно выразить форму этой связи с помощью регрессионного анализа.

Цель регрессионного анализа – с помощью уравнения регрессии предсказать ожидаемое среднее значение результирующей переменной.

Основные задачи регрессионного анализа следующие:

- определения вида и формы зависимости;
- оценка параметров уравнения регрессии;
- проверка значимости уравнения регрессии;

- проверка значимости отдельных коэффициентов уравнения;
- построение интервальных оценок коэффициентов;
- исследование характеристик точности модели;
- построение точечных и интервальных прогнозов результирующей переменной.

Как и корреляционный анализ, регрессионный анализ отражает только количественные зависимости между переменными. Причинно-следственные зависимости регрессионный анализ не отражает. Гипотезы о причинно-следственной связи переменных должны формулироваться и обосновываться исходя из теоретического анализа содержания изучаемого явления.

1. Основы корреляционного и регрессионного анализа.

Корреляция — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми).

Корреляционный анализ — метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными.

Ограничения корреляционного анализа:

1) Применение возможно при наличии достаточного количества наблюдений для изучения. На практике считается, что число наблюдений должно не менее чем в 56 раз превышать число факторов.
2) Необходимо, чтобы совокупность значений всех факторных и результативного признаков подчинялась многомерному нормальному распределению.
3) Исходная совокупность значений должна быть качественно однородной.
4) Сам по себе факт корреляционной зависимости не даёт основания утверждать, что одна из переменных предшествует или является причиной изменений, или то, что переменные вообще причинно связаны между собой, а не наблюдается действие третьего фактора.

Регрессионный анализ

❑ Оценка связи между двумя переменными (количественными – линейный регрессионный анализ, порядковыми тоже возможно, но точность анализа меньше)

- одна из переменных, x, называется независимой переменной, а другая, y, — зависимой. Набор значений y, соответствующих определенному значению x, обозначим y|x. среднее в точке x обозначим $\mu_{y|x}$

$$\mu_{y|x} = \alpha + \beta x.$$

Здесь α — значение y в точке x = 0 (коэффициент сдвига), β — коэффициент наклона

УСЛОВИЯ ПРИМЕНИМОСТИ

- Среднее значение $\mu_{y|x}$ линейно зависит от x.
- Для любого значения x значения y|x распределены нормально.
- Стандартное отклонение $\sigma_{y|x}$ одинаково при всех значениях x.

1 Регрессионный анализ

Функциональная зависимость может быть представлена в виде «ящика»: он преобразует вход $X = \{x_1, x_2, \ldots, x_N\}$, к выходу $Y = \{y_1, y_2, \ldots, y_N\}$

.

Функция ящика: одномерная («один вход» - «один выход»), или многомерная.

что известно об объекте:

все	структура	ничего

	белый	серый	черный
структура	+	+	-
колич. значения параметров	+	-	-

Определение F критерия Фишера

Так как в большинстве случаев уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос об адекватности построенного уравнения данным генеральной совокупности. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

где n – число наблюдений;

m – число факторов в уравнении регрессии.

Если в уравнении регрессии свободный член $a_0 = 0$, то числитель n-m-1 следует увеличить на 1, т.е. он будет равен n-m.

Определение коэффициента детерминации R^2

Для анализа общего качества уравнения линейной многофакторной регрессии используют множественный коэффициент детерминации R^2, называемый также квадратом коэффициента множественной корреляции R

$$R^2 = \frac{\sigma_F^2}{\sigma_y^2}$$

и определяет долю вариации результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную регрессионную модель.