Простая линейная регрессия и простой корреляционный анализ

Пары точек наносятся на координатную сетку. Из этого получают общее предварительное представление о рассеянии облака точек (1 - нет ошибок измерения, 2- есть ошибки измерения, 3 - присутствуют ошибки измерения и изменения, то есть отсутствует явно выраженный тренд). В случае 1 - функциональная зависимость, в случае 2,3 - стохастическая.

Функциональная зависимость - это закон, ставящий в соответствие каждому действительному числу X из множества

действительное число Y из множества

.

Стохастическая (случайная) зависимость между величинами X и Y - это зависимость, при которой строго определенному значению величины X может соответствовать множество значений величины Y. Зависимость носит вероятностный характер, то есть СВ Y принимает разные значения с некоторой вероятностью. Термин “стохастическая” связь впервые введен русскими статистиком А.А. Чупровым в 1926 г.

По отношению к событиям функциональная зависимость - причинна, то есть наступление одного события влечет за собой другое. Стохастическая связь при этом, наоборот, не является причинной. Одно событие изменяет вероятность наступления другого, однако не всегда вызывает его появление. Функциональная зависимость является предельным случаем стохастической - при наиболее тесной связи. Другой крайний случай - полная независимость величины Y от величины X. Таким образом стохастическая связь может быть более или менее тесной.

При оценивании стохастической зависимости различают корреляцию (существует ли взаимосвязь между переменными) и регрессию (какая зависимость).

Совокупность характеристик стохастической связи исследователи делят условно на две группы: в первую входят характеристики, позволяющие дифференцировать линейную и нелинейную зависимости, во вторую - характеристики степени тесноты связи (меры связи).

В [Мирский Г.Я. Характеристики стохастической взаимосвязи и их измерения. - М.:Энергоиздат, 1982] приводятся и подробно обсуждаются вероятностные характеристики, с помощью которых описывается стохастическая связь случайных процессов и случайных величин, а также способы их измерения, в том числе:

1) корреляционная функция и взаимная корреляционная функция;

2) нормированная и нормированная взаимная корреляционные функции;
3) коэффициент корреляции (полный, частный, сводный);
4) корреляционное отношение;
5) моментная функция и взаимная моментная функция;
6) функция регрессии;
7) функция когерентности;
8) условная функция распределения вероятностей;
9) условная плотность распределения вероятностей;
10) коэффициент коллигации и функция коллигации;
11) ранговый коэффициент корреляции Спирмэна и Кендалла и др.

Можно утверждать, что связанные случайные процессы не всегда являются взаимосвязанными. Так на изменение кардиограммы оказывают влияние колебания атмосферного давления, изменение состояния окружающей среды и т. д., хотя сами климатические условия от кардиограммы не зависят.

Наиболее часто используют такие характеристики стохастической связи как коэффициент и функция корреляции. Корреляционный анализ [Фестер Э., Ренц Б. Методы регрессионного и корреляционного анализа.- М.: Финансы и статистика, 1983], дающий хорошие результаты при линейных зависимостях, недостаточно эффективен, однако, в случае нелинейной связи. Поэтому представляется полезным исследовать коэффициент взаимосвязанности (коллигации), примененный академиком С.Н. Бернштейном для оценки степени зависимости двух случайных событий
и
.

Коэффициент коллигации служит характеристикой жесткости связи при любом ее характере - линейном и нелинейном, позволяет обнаруживать и достоверно констатировать независимость случайных величин.

Корреляционный анализ изучает на основании выборки стохастическую зависимость между случайными величинами X,Y.
Ковариация определяется как математическое ожидание произведения отклонений случайных величин:

,

Смешанный центральный момент второго порядка
Для устранения недостатка ковариации был введён **линейный коэффициент корреляции** (или **коэффициент корреляции**

Пирсона), который разработал Карл Пирсон в 90-х годах XIX века. Коэффициент корреляции рассчитывается по формуле:

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}.$$

где , — среднее значение выборок.

Коэффициент корреляции изменяется в пределах от минус единицы до плюс единицы

Функция от моментов второго порядка

называется коэффициентом корреляции

В математической статистике регрессионным анализом называют совокупность приемов для установления связей между независимой переменной *Y* и одной или несколькими переменными.

Регрессия - условное математическое ожидание случайной переменной *Y* при условии, что другая условная переменная *X* приняла значение *x*.

Моделью линейной регрессии является модель, в которой теоретическое среднее значение

наблюдаемой величины *y* является линейной комбинацией независимых переменных

$$y=a_0+a_1x_1+a_2x_2+... \qquad (1)$$

для случая, когда в модель включаются *k* переменных *x*.

Множители a_i, $i=1,2,...$ представляют собой параметры модели, значения которых должны быть установлены. Они называются коэффициентами регрессии, а a_0 называется свободным или постоянным членом.

Модель, более чем с одной переменной *x (1)* называется моделью множественной регрессии.