ПРОСТОЙ ПРИМЕР ЛИНЕЙНОЙ РЕГРЕССИИ

Из статьи Вы узнаете основы регрессионного анализа: как выбирают регрессионную модель, какие регрессионные модели бывают и для чего вообще нужна эта модель. Также, какие методы определения качества модели используют.

ПРОБЛЕМА РЕГРЕССИИ

В изучении любых реальных процессов, будь то варка макарон или анализ инвестиций, есть один общий принцип - они все зависят от каких-либо параметров. Вкус макарон зависит от температуры плиты, количества воды, соли, качества макарон и так далее, математически это обозначается так:

```
Вкус = f(температура, объём воды, соль, ...)
```

Итак, разберёмся с варкой порции макарон, у Вас набор случайных величин: температура плиты, объём воды, количество соли. Зададимся целью узнать, как количество воды влияет на вкус макарон.

ПОСТАНОВКА ЗАДАЧИ

Как определить влияние объёма воды на вкус макарон? Необходимо провести ряд экспериментов, в которых каждая варка макарон будет проводиться с разным объёмом воды, но остальные условия (температура и количество соли) будут фиксированы. Зададимся значениями температуры и количеством соли:

Температура	t=500°C
Количество соли	15 г

Таблица 1. *Фиксированные значения для эксперимента*

Начнём наши эксперименты для различных объёмов воды, возьмём от 500 мл до 2200 мл, и каждый раз будем пробовать макароны на вкус и запишем все наши результаты:

#	Объём воды	Оценка
1	500 мл	2

#	Объём воды	Оценка
2	600 мл	3
3	700 мл	4
4	800 мл	5
5	900 мл	6
6	1000 мл	8
7	1100 мл	11
8	1200 мл	12
9	1300 мл	14
10	1400 мл	19
11	1500 мл	21
12	1600 мл	25
13	1700 мл	30
14	1800 мл	34
15	1900 мл	50
16	2000 мл	51
17	2100 мл	62
18	2200 мл	78

Таблица 2. *Оценка вкуса макарон в зависимости от объёма воды*

ВЫЯВЛЕНИЕ ЗАВИСИМОСТИ

Итак, мы оцениваем вкус макарон в зависимости от объёма воды, математически мы изучаем функцию: Вкус = f(Объём). Весь регрессионный анализ заключается в процессе выявления функции f в данной зависимости.

В регрессионном анализе, функции (модели) делятся на два типа: линейные и нелинейные.

> *Линейная модель*
> $y = a + bx$
> *Нелинейная модель*
> $y = ab^x + c$

Для того, что бы построить **простую** регрессионную модель (функцию), необходимо набраться мужества и выдвинуть предположение, например:

> — Эта функция похожа на линейную!

Когда Вы выбрали регрессионную модель, Вы начинаете подбирать

коэффициенты, например, в линейной модели y=a+bx, необходимо подобрать коэффициенты a и b. Задача относительно не сложная, "a" - это первое значение, а "b" можно найти разницей последнего и первого значений. Провернув такую операцию с нашим примером, получим:

```
a = -20.5
b = 0.045
Вкус = -20.5 + 0.045x
```

Затабулируем значения нашей модели:

500 мл	**600 мл**	**700 мл**	**800 мл**	**900 мл**	**1000 мл**	**1100 мл**	**1200 мл**	**1300 мл**
2	6.5	11	15.5	20	24.5	29	33.5	38
1400 мл	**1500 мл**	**1600 мл**	**1700 мл**	**1800 мл**	**1900 мл**	**2000 мл**	**2100 мл**	**2200 мл**
42.5	47	51.5	56	60.5	65	69.5	74	78.5

Таблица 3. *Затабулированные значения регрессионной модели*

Вот, как это выглядит на графике:

График 1. *Линейная регрессионная модель и исходные данные*

ПОЛУЧЕНИЕ РЕЗУЛЬТАТА

С натяжечкой, конечно, похоже, но для математического вывода необходимо найти разброс значений модели и реальных значений. Эти значения - сумма квадратов отклонений и среднеквадратическая ошибка:

```
RSS (сумма квадратов отклонений) = (2 - 2)^2 + (6.5 - 3)^2 + ... + (78.5 - 78)^2 = 6022.25
MSE (среднее квадратическое отклонение) = √RSS = 77.6

S (дисперсия) = 18.29
```

Что делать с этой регрессионной моделью? Регрессионная модель позволяет предсказать, а что будет, например, если мы возьмём 2300 мл, 2400 мл и т.д. не проводя при этом сам эксперимент:

$$Вкус_{2300\ мл} = -20.5 + 0.045 \cdot 2300 = 83$$

$Вкус_{2400\ мл} = -20.5 + 0.045 \cdot 2400 = 87.5$

И, разумеется, мы можем узнать сколько нужно воды для идеальных макарон:

$Вода_{идеальные\ макароны} = (100-20.5) / 0.045 = 2678\ мл$

МИНИМИЗИРУЕМ ОШИБКУ

Итак, с нами наша модель y = a + bx и реальные значения функции, разница между ними - это и есть ошибка, которую мы допускаем в каждом эксперименте. Значит, мы можем построить функцию ошибки, а если у нас есть функция, то мы всегда можем найти её минимум. Этим мы и займёмся, нахождением минимума функции ошибки.

Ошибка - это разница между реальным значением и смоделированным, поскольку эта разница может быть как положительной, так и отрицательной, необходимо использовать модуль разницы, что проще всего сделать возведя ошибку в квадрат, а затем извлечь корень. Значит, наша ошибка на каждом известном результате:

Y_o - значение из наблюдений (observation), Y_m - значение из модели (model)
$e = (Y_o - Y_m)^2 = (Y_o - a - bx)^2$
Суммарная ошибка
$S = \Sigma e = \Sigma(Y_o - a - bx)^2$

Функция S - это функция ошибки, которую необходимо минимизировать, она зависит от параметров a и b. Для нахождения минимума функции воспользуемся простым методом - найдём производные по параметрам a и b (здесь мы опустим сложные методы поиска минимума функции):

Производные функции ошибки по параметрам a и b:
$dS/da = \Sigma 2(a+bx-y)$
$dS/db = \Sigma 2(a+bx-y)x$
Условие минимума функции:
$\Sigma 2(a+bx-y) = 0$
$\Sigma 2(a+bx-y)x = 0$

Упростим, сократим на 2 и разложим скобки (n-количество наблюдений):

$na + b\Sigma x = b\Sigma y$

$a\Sigma x + b\Sigma x^2 = \Sigma xy$

Найдём решение:

$\Sigma x = 24\ 300$

$\Sigma x^2 = 37\ 650\ 000$

$\Sigma y = 435$

$\Sigma xy = 776\ 800$

$18 \cdot a + 24300 \cdot b = 435$

$24300 \cdot a + 37650000 \cdot b = 776800$

$-3589 \cdot a = 102819 \Rightarrow a = -29$

$b = 0.039$

Попробуем нашу новую модель в действии:

График 3. *Линейная регрессионная модель урегулированная методом наименьших квадратов*

RSS (сумма квадратов отклонений) $= (-9.5 - 2)^2 + (-5.6 - 3)^2 + \ldots + (56.8 - 78)^2 = 1203.65$

MSE (среднее квадратическое отклонение) $= \sqrt{RSS} = 34.69$

S (дисперсия) $= 8.18$

$Вкус_{2300\ мл} = -29 + 0.039 \cdot 2300 = 60.7$

$Вкус_{2400\ мл} = -29 + 0.039 \cdot 2400 = 64.6$

Как Вы, наверное, заметили, предсказания по нашей первой модели ближе к правде, нежели модели отрегулированной. Почему? Потому что модель была выбрана неверно, график функции больше похож на экспоненту и даже исходя из знания процесса ясно, что линейной зависимости здесь не место. Но это был всего лишь пример линейной регрессионной модели, о более сложных моделях и о способе выбора модели читайте в следующих статьях.

УДК: ГРНТИ:
Автор статьи: *Телятников З.А.*
Дата написания статьи: *07.06.2017*
Дата редакции статьи: *08.06.2017*
Адрес статьи в интернете:
http://k-tree.ru/articles/statistika/prognozirovanie/prostoi_primer_lineinoi_regressii

Дата формирования документа: *20.11.2017 18:22*