10. Обработка эксперимнт данных. Линейный регрисеонный анализ. коэф-ты ур-я регресии. Коэф-ты корреляции. Стандартные отклонения коэф регрессии. Проверка значимости

Моделирование числа поступивших в университет для лучшего понимания факторов, удерживающих детей в том же учебном заведении. 2. Моделирование потоков миграции в зависимости от таких факторов как средний уровень зарплат, наличие медицинских, школьных учреждений, географическое положение… 3. Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д., 4. Моделирование потерь от пожаров как функции от таких переменных как количество пожарных станций, время обработки вызова, или цена собственности. Суть регрессионного анализа заключается в нахождении наиболее важных факторов, которые влияют на зависимую переменную.

Уравнение регрессии. Это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$$

Зависимая переменная (Y) — это переменная, описывающая процесс, который мы пытаемся предсказать или понять. Независимые переменные (X) это переменные, используемые для моделирования или прогнозирования значений зависимых переменных. В уравнении регрессии они располагаются справа от знака равенства и часто называются объяснительными переменными. Зависимая переменная - это функция независимых переменных. Коэффициенты регрессии (β) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой. Невязки. Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки ε.

Различают линейные и нелинейные регрессии.

Линейная регрессия: $y = a + b \cdot x + \varepsilon$

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака у от теоретических минимальна.

$$\begin{cases} Y^1 = \alpha + \beta_1 X_1^1 + \ldots + \beta_k X_k^1 + u^1, \\ Y^2 = \alpha + \beta_1 X_1^2 + \ldots + \beta_k X_k^2 + u^2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Y^n = \alpha + \beta_1 X_1^n + \ldots + \beta_k X_k^n + u^n, \end{cases}$$

$YX^{-1} = XX^{-1}B$ ⇒ $YX^{-1} = B$

$XX^{-1} = 1$

Таким образом, мы получим в явной форме набор уравнений на компоненты вектора B, то есть наше искомое решение регрессионной задачи.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \cdot \begin{bmatrix} x_{01} & x_{11} & \ldots & x_{k1} \\ x_{02} & x_{12} & \ldots & x_{k2} \\ \vdots & \vdots & \ldots & \vdots \\ x_{0N} & x_{1N} & \ldots & x_{kN} \end{bmatrix}^{-1} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

В общем случае коэффициент регрессии *k* показывает, как в среднем изменится *результативный признак* (*Y*), если *факторный признак* (*X*) увеличится на единицу .
Y = 87610 + 2984 *X* ; ***X*** *– число рабочих,* ***Y*** *– объем годового производства (руб.).*
Пример интерпретации коэффициента регрессии
• В уравнении *Y* = 87610 + **2984** *X* ; коэффициент регрессии равен +2984.
Что это означает?
•В данном случае смысл коэффициента регрессии состоит в том, что увеличение *числа рабочих* на 1 чел. приводит в среднем к увеличению объема годового *производства* на 2984 руб.

Свойства коэффициента регрессии

• Коэффициент регрессии может принимать любые значения.
• Коэффициент регрессии *не симметричен* , т.е. изменяется, если *X* и *Y* поменять местами.
• *Единицей измерения* коэффициента регрессии является отношение единицы измерения *Y* к единице измерения *X* : ([*Y*] / [*X*]).
• Коэффициент регрессии *изменяется при изменении единиц измерения* X и Y .
• Поскольку результативный признак *Y* измеряется в рублях, а факторный признак *X* в количестве рабочих (чел.), то коэффициент регрессии измеряется *в рублях на человека* (руб. / чел.)

Коэффициент детерминации рассматривают, как правило, в качестве основного показателя, отражающего меру качества регрессионной модели, описывающей связь между зависимой и независимыми переменными модели. Коэффициент детерминации показывает, какая доля вариации объясняемой переменной *y* учтена в модели и обусловлена влиянием на нее факторов, включенных в модель:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

где – y_i значения наблюдаемой переменной, $\bar{y}$ – среднее значение по наблюдаемым данным, $\hat{y}_i$ – модельные значения, построенные по оцененным параметрам.

1 Обработка экспериментальных данных

Как правило, в результате эксперимента получают калибровочные данные, которые можно представить в виде уравнения

$$y = a_0 + a_1 x + a_2 x^2 + \ldots,$$

где x – независимая переменная;
y – зависимая переменная;
a_0, a_1,a_2 – коэффициенты уравнения.

Обычно для заданной таблицы экспериментальных данных возникает задача определения такого набора коэффициентов, при котором уравнение наилучшим образом приближает эти экспериментальные данные.

Каждая ордината y_i в эксперименте определяется с некоторой ошибкой, которую можно характеризовать, например, **средним квадратичным отклонением,** т.е.

$$y_i \pm S_i,$$

где

$$S_i = \sqrt{\frac{\sum_{i=1}^{n}(\bar{y} - y_i)^2}{n-1}},$$

где $\bar{y}$ – среднее арифметическое n экспериментальных значений:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Еще линейную регрессию часто называют методом наименьших квадратов, поскольку коэффициенты a и b вычисляются из условия минимизации суммы квадратов ошибок $|b + ax_i - y_i|$.

Для нахождения коэффициентов уравнения регрессии используются следующие формулы:

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})},$$

$$b = \bar{y} - a\bar{x},$$

где $\bar{x}$ - среднее арифметическое x:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Стандартные отклонения коэффициентов уравнения регрессии находят по формуле:

$$S_a = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}},$$

$$S_b = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}.$$

Рассеяние результатов относительно прямой оценивают с помощью дисперсии S_y^2, которая находится по формуле:

$$S_y^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$

корень квадратный из которой определяет стандартное отклонение точек от найденной зависимости.

При проведении некоторых химико-технологических исследований возникает необходимость оценить характер и степень зависимости одной экспериментальной величины от другой или нескольких исследуемых величин, т.е. с точки зрения математической статистики следует установить корреляцию между случайными величинами. Чаще всего ищут линейную зависимость. С этой целью используют безразмерный коэффициент корреляции ρ, рассматриваемый как мера отклонения зависимости случайных величин от линейной.

Для нахождения коэффициента корреляции используется выражение:

$$\rho = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}},$$

Если коэффициент корреляции равен по модулю единице, то между случайными величинами существует линейная зависимость. Если же он равен нулю, то случайные