ЛАБОРАТОРНАЯ РАБОТА №2

«Множественная линейная регрессия»

2.1 Описание метода.

Определение: Уравнение, связывающее один из признаков зависимостью от других признаков, называется уравнением ***регрессии.*** Уравнение регрессии зависит от неизвестных параметров.

Классический регрессионный анализ занимается моделями, линейными по параметрам.

Уравнение линейной множественной регрессии:

$$y = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_m x^m + \varepsilon \qquad (*)$$

это векторное равенство, где $x^i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ni} \end{pmatrix}$ — вектор независимых переменных, а $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ — вектор неизвестных параметров, ε — вектор, играющий роль случайной помехи.

Векторное равенство (*) можно записать в виде:

$$y_k = \alpha_1 x_{k1} + \alpha_2 x_{k2} + \cdots + \alpha_m x_{km} + \varepsilon_k,\ k = \overline{1, N}.$$

Здесь ε — случайная компонента, комплексно характеризующая эффект неучтенных признаков.

Введем в рассмотрение матрицу *X*:

$$X = \begin{pmatrix} x_{11} & \ldots & x_{1m} \\ \ldots & \ldots & \ldots \\ x_{N1} & \ldots & x_{Nm} \end{pmatrix}.$$

Тогда можем записать уравнение линейной множественной регрессии в матричном виде:

$$y = X\alpha + \varepsilon \qquad (**).$$

2.2. Постулаты (предположения) регрессионного анализа.

Так как в уравнении регрессии фигурируют матрица данных *X*, вектор неизвестных параметров α и вектор случайной помехи ε, то предположения регрессионного анализа касаются этих трех элементов.

1) На вектор α ограничений не наложено, $\alpha \in R^m$;

2) ε — случайный вектор, следовательно, вектор y - случайный;

3) Математическое ожидание всех компонент вектора ε равно нулю: $M(\varepsilon_k) = 0,\ k = \overline{1,N}$;

4) Ковариация между ε_k и ε_j:

$$\operatorname{cov}(\varepsilon_k, \varepsilon_j) = \begin{cases} 0, k \neq j \\ \sigma^2, k = j \end{cases},\ k, j = \overline{1,N}.$$

То есть у различных объектов случайные помехи не коррелированны, а дисперсия вектора ε конечна и одинакова для всех наблюдений (условия проведения наблюдений одинаковы для всех объектов).

5) Матрица *X* детерминирована (не случайна), то есть значения независимых признаков известны исследователю точно.

6) Ранг матрицы *X* равен m, то есть в матрице X имеется m линейно независимых строк или столбцов.

2.3. Суть МНК.

Суть МНК состоит в следующем: параметры выбираются из условия минимума суммы квадратов отклонений фактических значений от расчетных. Сумму квадратов отклонений фактических значений от расчетных обозначают $Q(\alpha)$.

$$Q(\alpha) = \sum_{k=1}^{n} \left(y_k - \alpha_1 x_{k1} - \alpha_2 x_{k2} \ - \ldots - \alpha_m x_{km}\right)^2$$

Взяв производную от $Q(\alpha)$ по вектору α и приравняв ее нулю, получили уравнение, из которого выразили α. Пусть *a* - МНК-оценка вектора α. Тогда

$$a = \left(X^T X\right)^{-1} X^T y.$$

2.4. Уравнение регрессии со свободным членом.

В силу 3-го постулата регрессионного анализа считается, что эффект неучтенных признаков в среднем равен 0. Это предположение на практике малоправдоподобно. Чаще эффект неучтенных факторов не 0, тогда вместо постулата 3 вводят постулат 3′: $M(\varepsilon_k) = \alpha_{m+1}$=const, где $\alpha_{m+1} \in$ R.

Тогда уравнение регрессии будет иметь вид:

$y_k = \alpha_1 x_{k1} + \alpha_2 x_{k2} + \cdots + \alpha_m x_{km} + \varepsilon_k'$, где $\varepsilon_k' = \varepsilon_k - \alpha_{m+1}$, k=1,...,N

Тогда $M(\varepsilon_k') = 0$

Мы оказались в условиях предыдущей системы постулатов, поэтому далее будем считать, что уравнение регрессии имеет вид:

$y_k = \alpha_1 x_{k1} + \ldots + \alpha_m x_{km} + \alpha_{m+1} x_{k,m+1} + \varepsilon_k'$, где $X^{m+1} = \begin{bmatrix} 1 \\ \ldots \\ 1 \end{bmatrix} = I$

$\alpha = \begin{bmatrix} \alpha_1 \\ ... \\ \alpha_{m+1} \end{bmatrix}$, а X(N×(m+1))

2.5. Среднее значение расчетных и фактических данных зависимых переменных.

Уравнение регрессии имеет вид: $y_k = \alpha_1 x_{k1} + ... + \alpha_m x_{km} + \varepsilon_k$, где x_{km}– фиктивная переменная.

Вычислим МНК оценку неизвестных параметров:

а- оценка α $a = \left(X^T X\right)^{-1} X^T y$.

Находим вектор расчетных значений зависимой пременной:

$$\hat{y} = Xa$$

Тогда y=Xa+e, где e= $y - \hat{y}$, где вектор $e = \begin{bmatrix} e_1 \\ ... \\ e_n \end{bmatrix}$.

Вектор e называется вектором оценочных отклонений.

МНК оценка удовлетворяет уравнению:

$-X^T y + X^T Xa = 0 \Rightarrow$ $X^T(y - Xa) = 0 \Rightarrow$ $X^T(y - \hat{y}) = 0 \Rightarrow X^T e = 0.$

Рассмотрим последнюю строку матрицы X^T. Это единицы

$I^T e = 0,$

$\sum_{k=1}^{N} e_k = 0, \ \frac{1}{N}\sum_{k=1}^{N} e_k = 0 \quad \Rightarrow \bar{e} = 0$

Вернемся к равенству $y = \hat{y} + e \Rightarrow y_k = \hat{y}_k + e_k$, просуммируем по k и разделим на N:

$\frac{1}{N}\sum_{k=1}^{N} y_k = \frac{1}{N}\sum_{k=1}^{N} \hat{y}_k + \frac{1}{N}\sum_{k=1}^{N} e_k \Rightarrow \bar{y} = \bar{\hat{y}} + \bar{e}, \ \bar{e} = 0 \Rightarrow \bar{y} = \bar{\hat{y}}.$

Т.е. среднее расчетное значение и среднее фактическое значение совпадают.

Коэффициент детерминации

$$R = 1 - \frac{\sum_{k=1}^{N} e_k^2}{\sum_{k=1}^{N} \left(y_k - \bar{y}\right)^2}$$

Коэффициент детерминации изменяется в пределах от 0 до 1. Он показывает, как велика доля объясненной дисперсии в общей дисперсии, какая часть общей дисперсии может быть объяснена зависимостью переменной y от переменных $x_1, x_2, \ldots, x_m$.

2.6. Задание и порядок выполнения работы.

Дан вектор y длины N и матрица X размера N×m. Предполагается, что между переменной y и переменными X^1, X^2, ..., X^m существует линейная зависимость (X^1, X^2, ..., X^m — столбцы матрицы X):

$$y = a_1 X^1 + a_2 X^2 + ... + a_m X^m + \varepsilon \text{, где } M(\varepsilon) = 0.$$

Найти МНК-оценку вектора коэффициентов линейной множественной регрессии, пользуясь формулой $a = (X^T X)^{-1} X^T y$.

1) В качестве вектора y взять один из столбцов матрицы Z (из лабораторной работы «Корреляционный анализ»). Тогда остальные столбцы матрицы Z(может быть не все) составят матрицу X.

Замечание. 1.Если модуль коэффициента корреляции между целевым признаком y и некоторым независимым признаком $x^i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ni} \end{pmatrix}$ ***меньше 0.3, этот независимый признак удаляем из матрицы X. Если осталось слишком мало столбцов (1 или 2) можно собрать в матрицу X те признаки, у которых с признаком y по результатам лаб. работы 1 есть связь (справедлива гипотеза H_1), если в этом случае в матрице X останется больше столбцов.***

2. Проблемой множественной линейной регрессии является мультиколлинеарность. Под мультиколлинеарностью понимается высокая взаимная коррелированность объясняющих(независимых) переменных. Для устранения или уменьшения мультиколлинеарности используется ряд методов. Самый простой состоит в том, что из двух независимых переменных, имеющих высокий коэффициент корреляции(больше по модулю 0,8), одну переменную исключают из рассмотрения. При этом, какую переменную оставить, а какую исключить решают в первую очередь из практических соображений.

2) Рассмотреть уравнение регрессии со свободным членом, для этого нужно изменить матрицу X (см. пункт 2.4.)

3) Составить программу для нахождения МНК-оценки вектора коэффициентов уравнения линейной множественной регрессии по формуле $a = (X^T X)^{-1} X^T y$. Предусмотреть в программе

проверку равенства средних значений расчетных и фактических данных зависимых переменных и вычисление коэффициента детерминации.

4) Составить тестовый пример. Для тестового примера взять матрицу размером 5x2 и вектор y длины 5. Выполнить вычисление по формуле $a = \left(X^T X\right)^{-1} X^T y$ вручную

5) Проверить составленную программу по тестовому примеру и найти МНК оценку для своих исходных данных.