Тема 2. Парная регрессия и корреляция.

Вопросы

- ✔ ***Статистическая зависимость (независимость) случайных переменных. Ковариация.***
- ✔ ***Анализ линейной статистической связи экономических данных, корреляция; вычисление коэффициентов корреляции.***
- ✔ ***Линейная модель парной регрессии.***
- ✔ ***Оценка параметров модели с помощью метода наименьших квадратов (МНК).***
- ✔ ***Оценка существенности параметров линейной регрессии.***
- ✔ ***Интервалы прогноза по линейному уравнению регрессии.***
- ✔ ***Нелинейные модели и их линеаризация***

Категории зависимости:
1) функциональные;
2) корреляционные.

Корреляционные связи:

1) между изменением факторного и результативного признака нет полного соответствия,
2) воздействие отдельных факторов проявляется лишь в среднем при массовом наблюдении фактических данных.
3) Одновременное воздействие на изучаемый признак большого количества самых разнообразных факторов приводит к тому, что одному и тому же значению признака-фактора соответствует целое распределение значений результативного признака, поскольку в каждом конкретном случае прочие факторные признаки могут изменять силу и направленность своего воздействия.

Функциональные связи характеризуются:

1) полным соответствием между изменением факторного признака и изменением результативной величины
2) каждому значению признака-фактора соответствуют вполне определенные значения результативного признака.
3) Функциональная зависимость может связывать результативный признак с одним или несколькими факторными признаками.

Задачи корреляционного анализа:

1) выявлении взаимосвязи между случайными переменными путем точечной и интервальной оценки парных (частных) коэффициентов корреляции, вычисления и проверки значимости множественных коэффициентов корреляции и детерминации.
2) отбор факторов, оказывающих наиболее существенное влияние на результативный признак, на основании измерения степени связи между ними;
3) обнаружение ранее неизвестных причинных связей.

При проведении корреляционного анализа вся совокупность данных рассматривается как множество переменных (факторов), каждая из которых содержит ***n* –наблюдений**.

При изучении взаимосвязи между двумя факторами их, как правило, обозначают **X=** $(x_1, x_2, \ldots, x_n)$ **и Y=** $(y_1, y_2, \ldots, y_n)$

Ковариация - это статистическая мера взаимодействия двух переменных.

Ковариация между двумя переменными X и У рассчитывается следующим образом:

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

где - фактические значения случайных переменных X и Y,

или $(x_1,y_1),\ (x_2,y_2),\ldots,(x_n,y_n)$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \qquad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Вычисление коэффициента парной корреляции.

Коэффициент парной корреляции

Для двух переменных X и У *коэффициент парной корреляции* определяется следующим образом:

$$r_{x,y}=\frac{Cov(X,Y)}{S_x\cdot S_y} \qquad \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_x S_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

Где $S_x^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2$

и $S_y^2=\frac{1}{n-1}\sum(y_i-\bar{y})^2$

- оценки дисперсий величин

Дисперсия (оценка дисперсии)

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

характеризует степень разброса значений $x_1, x_2, x_3, \ldots, x_n$ ($y_1, y_2, y_3, \ldots, y_n$) вокруг своего среднего $\bar{x}$ ($\bar{y}$, соответственно), или вариабельность (изменчивость) этих переменных на множестве наблюдений.

В общем случае для получения несмещенной оценки дисперсии сумму квадратов следует делить на число степеней свободы оценки (*n*–*p*), где *n* - объем выборки, *p* - число наложенных на выборку связей. В данном случае *p* = 1, т.к. выборка уже использовалась один раз для определения среднего *X*, поэтому число наложенных связей равно единице, а число степеней свободы оценки (т.е. число независимых элементов выборки) равно (n –1).

Среднеквадратическое отклонение или стандартное отклонение, или стандартная ошибка переменной X (переменной Y)

$$S_x = \sqrt{S_x^2}$$

Оценка значимости коэффициента корреляции при малых объемах выборки выполняется с использованием ***t - критерия Стьюдента***. При этом фактическое (наблюдаемое) значение этого критерия определяется по формуле:

$$t_{набл} = \sqrt{\frac{r_{y,x}^2}{1 - r_{y,x}^2}(n-2)}$$

Парная линейная регрессия

Парная регрессия – это уравнение связи двух переменных x и y

$$y = \hat{f}(x)$$

где x - независимая, объясняющая переменная (признак-фактор),

y - зависимая переменная (результативный признак).

Замечание. Число наблюдений должно в 7-8 раз превышать число рассчитываемых параметров при переменной .

Пусть имеется набор значений двух переменных: $Y = (y_1, y_2, y_3, \ldots, y_n)$ - объясняемая переменная и $X = (x_1, x_2, x_3, \ldots, x_n)$ - объясняющая переменная, каждая из которых содержит *n* наблюдений, между которыми теоретически существует некоторая линейная зависимость

$$Y = f(X) = f(x_1, x_2, \ldots x_n) = \alpha + \beta x$$

Учитывая возможные отклонения, линейное уравнение связи двух переменных (парную регрессию) представим в виде:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

где - α постоянная величина (или свободный член уравнения),

- β коэффициент регрессии, определяющий наклон линии, вдоль которой рассеяны данные наблюдений. Это показатель, характеризующий изменение переменной y_i, при изменении значения x_i на единицу.

Если $\beta > 0$ - переменные x_i и y_i положительно коррелированные, если $\beta < 0$ – отрицательно коррелированны; ε_i - случайная переменная, или случайная составляющая, или остаток, или возмущение. Она отражает тот факт, что изменение y_i будет неточно описываться изменением X – присутствуют другие факторы, неучтенные в данной модели.

Таким образом, в уравнении (2) значение каждого наблюдения y_i представлено как сумма двух частей — систематической $\alpha + \beta x_i$ и случайной ε_i

$\hat{y}_i = \alpha + \beta x_i$ таким образом

$$Y = \hat{Y} + \varepsilon$$

Предпосылки метода наименьших квадратов.

1) Математическое ожидание случайной составляющей в любом наблюдении должно быть равно нулю.

$$M(\varepsilon_i) = 0$$

2) Второе условие состоит в том, что в модели (2) возмущение ε_i (или зависимая переменная y_i) есть величина случайная, а объясняющая переменная x_i- величина неслучайная.

3) Третье условие предполагает отсутствие систематической связи между значениями случайной составляющей в любых двух наблюдениях.

$$M(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

4) Дисперсия случайной составляющей должна быть постоянна для всех наблюдений.

5) Предположение о нормальности

Свойства оценок МНК.

1. ***Несмещенность оценки*** означает, что математическое ожидание остатков равно нулю.
2. ***Оценки считаются эффективными***, если они характеризуются наименьшей дисперсией.
3. ***Состоятельность оценок*** характеризует увеличение их точности с увеличением объема выборки

Оценка параметров регрессионного уравнения

МНК минимизирует сумму квадратов отклонения наблюдаемых значений y_i от модельных значений $\hat{y}_i$.

Согласно принципу метода наименьших квадратов, оценки $\hat{\alpha}$ и $\hat{\beta}$ находятся путем минимизации суммы квадратов:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

В результате применения МНК получаем формулы для вычисления параметров модели парной регрессии.

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \qquad (3)$$

Такое решение может существовать только при выполнении условия

$$\sum_{i=1}^{n}(x_i-\bar{x})^2\neq 0,$$

что р отличию от нуля определителя системы нормальных уравнений. Действительно, этот определитель равен

$$n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2=n\left[\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right]=n\sum_{i=1}^{n}(x_i-\bar{x})^2 .$$

Послед ***мости*** *модели наблюдений*

$$y_i=(\alpha+\beta\cdot x_i)+\varepsilon_i,\ i=1,\dots,n$$

о не все значения *совпадают между собой.* $x_1,\dots,x_n$

точки $(x_i,y_i),\ i=1,\dots,n$

лежат на одной вертикальной прямой

Оценки *и* *называют* ***оценками наименьших квадратов****. Обратим внимание на полученное выражение для параметра* $\hat{\beta}$ *. В это выражение входят суммы квадратов, участвовавшие ранее в определении выборочной дисперсии* $\hat{\alpha}$

$$S_x^2=Var(x)=\sum_{i=1}^{n}(x_i-\bar{x})^2\Big/(n-1)$$

ии

терминах параметр β *можно*

$$Cov(x,y)=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\Big/(n-1),$$

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\Big/(n-1)}{S_x^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} =$$

$$= r_{x,y}\cdot\frac{S_y}{S_x} = \frac{\overline{y\cdot x}-\bar{y}\cdot\bar{x}}{\overline{x^2}-\bar{x}^2} = \frac{\sum_{i=1}^{n} y_i\cdot x_i - n\cdot\bar{y}\cdot\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2}$$

Оценка качества уравнения регрессии

После построения уравнения регрессии мы можем разбить значение **Y**, в каждом наблюдении на две составляющих - $\hat{y}_i$ и e_i.

$$y_i = \hat{y}_i + e_i$$

Остаток представляет собой отклонение фактического значения зависимой переменной от значения данной переменной, полученное расчетным путем:

$$e_i = y_i - \hat{y}_i \quad (i = \overline{1:n}).$$

$$e_i \neq 0$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \quad (4).$$

Где $\hat{y}_i$ - значения *y*, вычисленные по модели $\hat{y}_i=\alpha+\beta x_i$

Разделив правую и левую часть (4) на $\sum_{i=1}^{n}(y_i-\bar{y})^2$

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

получим

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

Коэффициент детерминации

$$R^2 = \frac{\text{объясняемая сумма квадратов}}{\text{общая сумма квадратов}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Коэффициент детерминации ***показывает долю вариации результативного признака, находящегося под воздействием изучаемых факторов, т. е. определяет, какая доля вариации признака Y учтена в модели и обусловлена влиянием на него факторов.***

для оценки качества регрессионных моделей целесообразно использовать **среднюю ошибку аппроксимации**:

$$E = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \times 100\%$$

Для **проверки значимости модели** регрессии используется **F-критерий Фишера**, вычисляемый как отношение дисперсии исходного ряда и несмещенной дисперсии остаточной компоненты.

Если расчетное значение с $N_1 = k$ и $N_2 = (n - k - 1)$ степенями свободы, где k – количество факторов, включенных в модель, больше табличного при заданном уровне значимости, то **модель считается значимой.**

Для модели парной регрессии:

$$F = \frac{R^2}{1-R^2} \cdot (n-2) = \frac{r_{y,x}^2}{1-r_{y,x}^2} \cdot (n-2)$$

В качестве **меры точности** применяют несмещенную оценку дис персии остаточной компоненты, которая представляет собой отношение суммы квадратов уровней остаточной компоненты к величине (n- k -1), где k – количество факторов, включенных в модель. Квадратный корень из этой величины (S_ε) называется **стандартной ошибкой**

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}}$$

Для модели парной регрессии

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - 2}}$$

Прогнозирование с применением уравнения регрессии

Прогнозируемое значение переменной y получается при подстановке в уравнение регрессии ожидаемой величины фактора x

$$\hat{y}_{прогн} = \hat{\alpha} + \hat{\beta} \cdot x_{прогн}$$

Доверительные интервалы, зависят от следующих параметров:

- стандартной ошибки,
- удаления $x_{прогн}$ от своего среднего значения $\bar{x}$
- количества наблюдений n
- и уровня значимости прогноза α.

В частности, для прогноза будущие значения $y_{прогн}$ с вероятностью (1 - α) попадут в интервал $y_{прогн} \in$

$$\left[\hat{y}_{прогн} - S_e \cdot t_\alpha \cdot \sqrt{1 + 1/n + \frac{(x_{прогн} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}};\ \hat{y}_{прогн} + S_e \cdot t_\alpha \cdot \sqrt{1 + 1/n + \frac{(x_{прогн} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right]$$

Нелинейные модели и их линеаризация

Задача построения нелинейной модели регрессии состоит в следующем:
Задана нелинейная спецификация модели

$$y = f(x,a,b,\varepsilon),$$

где
y - зависимая, объясняемая переменная;
x - независимая, объясняющая переменная;
a,b - параметры модели, для которых должны быть получены оценки;
ε - аддитивный или мультипликативный случайный фактор.

Требуется

1. Преобразовать исходные данные $x \rightarrow x^*$, $y \rightarrow y^*$ так, чтобы спецификация модифицированной регрессии была линейной:

$$y^* = a^* + b^* x^*$$

2. Методом наименьших квадратов получить оценки параметров a^*, b^*.
3. По оценкам a^*, b^* вычислить искомые оценки параметров a, b исходной регрессии.

Способы преобразования данных и вычисления параметров *a, b* по оценкам *a*, b** :

Исходная спецификация	Преобразование	Преобразование	Вычисление *b* по *b**	Вычисление *a* по *a**
	$x \to x^*$	$y \to y^*$		
$y=a+\frac{b}{x}+\varepsilon$	$x^*=\frac{1}{x}$	$y^*=y$	$b=b^*$	$a=a^*$
$y=\frac{1}{a+bx+\varepsilon}$	$x^*=x$	$y^*=\frac{1}{y}$	$b=b^*$	$a=a^*$
$y=\frac{x}{a+bx+x\varepsilon}$	$x^*=\frac{1}{x}$	$y^*=\frac{1}{y}$	$b=a^*$	$a=b^*$
$y=ae^{bx+\varepsilon}$	$x^*=x$	$y^*=\ln y$	$b=b^*$	$a=e^{a^*}$
$y=ae^{\frac{b}{x}+\varepsilon}$	$x^*=\frac{1}{x}$	$y^*=\ln y$	$b=b^*$	$a=e^{a^*}$
$y=\frac{1}{a+be^{-x}+\varepsilon}$	$x^*=e^{-x}$	$y^*=\frac{1}{y}$	$b=b^*$	$a=a^*$
$y=ax^be^{\varepsilon}$	$x^*=\ln x$	$y^*=\ln y$	$b=b^*$	$a=e^{a^*}$