ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ОРЕНБУРГСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Б1.Б.10 Эконометрика

Специальность 38.05.01 Экономическая безопасность

Специализация Экономико-правовое обеспечение экономической безопасности

Форма обучения заочная

1. Конспект лекций

1.1 Лекция №1 (1 час)

Тема: Парная линейная и нелинейная регрессия

1.1.1. Вопросы лекции:

1. Метод наименьших квадратов для построения линейной модели
2. Линейный коэффициент корреляции. Коэффициент детерминации
3. Проверка значимости регрессионной модели, проверка качества параметров уравнения и построение доверительных интервалов коэффициентов регрессии
4. Интерпретация уравнения регрессии
5. Средняя ошибка аппроксимации
6. Виды нелинейных зависимостей, поддающиеся линеаризации

1.1.2. Краткое содержание вопросов

1. Метод наименьших квадратов для построения линейной модели

Классический подход к оцениванию параметров линейной регрессии основан на *методе наименьших квадратов* (МНК), разработанный Гауссом.

Метод наименьших квадратов позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных (теоретических) минимальна:

$$S = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2 \longrightarrow \min \quad \text{т.е} \quad \sum \varepsilon_i^2 \to \min$$

Другими словами, из множества линий линия регрессии выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной.

Для линейной парной зависимости:

$$S = \sum_{i=1}^{n} \left[y_i - \left(a + b \cdot x\right)\right]^2 \longrightarrow \min .$$

2. Линейный коэффициент корреляции. Коэффициент детерминации

Уравнение регрессии всегда дополняется показателем тесноты связи. Для линейной регрессии таким показателем является **линейный коэффициент корреляции**.

Был впервые введен Карлом Пирсоном (1857 - 1936).

В теории разработаны и на практике применяются различные модификации формул расчета:

$$r = b\frac{\sigma_x}{\sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y},$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{\sum x_i \cdot f_i}{\sum f_i}; \quad \bar{y} = \frac{\sum y_i}{n} = \frac{\sum y_i \cdot f_i}{\sum f_i};$$

$$\overline{x \cdot y} = \frac{\sum x_i \cdot y_i}{n};$$

$$\sigma_x = \sqrt{\frac{\sum \left(x_i - \bar{x}\right) \cdot f_i}{\sum f_i}} = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n}} = \sqrt{\frac{\sum x_i^2}{n} - \left(\bar{x}\right)^2}$$

$$\sigma_y = \sqrt{\frac{\sum \left(y_i - \bar{y}\right) \cdot f_i}{\sum f_i}} = \sqrt{\frac{\sum \left(y_i - \bar{y}\right)^2}{n}} = \sqrt{\frac{\sum y_i^2}{n} - \left(\bar{y}\right)^2}$$

Линейный коэффициент корреляции находится в пределах:

$$-1 \le r \le 1$$

Чем ближе r по абсолютной величине к 1, тем теснее связь между признаками.

Знак указывает на направление связи

Если $|r| \ge 0{,}7$, то считают связь сильной;

$0{,}5 \le |r| \le 0{,}7$ – средней тесноты;

$|r| < 0{,}5$ – слабой.

Квадрат линейного коэффициента корреляции называется ***коэффициентом детерминации***. Этот коэффициент характеризует долю общей вариации результативного признака, которая объяснятся вариацией факторного признака. (Например, рассмотрели группировку производительности труда по квалификации рабочих и получили r^2=45,4%. Фактор квалификации рабочих объясняет 45,4% вариации производительности труда, а неучтенные факторы – 54,6%.).

3. Проверка значимости регрессионной модели, проверка качества параметров уравнения и построение доверительных интервалов коэффициентов регрессии

Оценка значимости параметров уравнения регрессии дается с помощью *t*-критерия Стьюдента. Выдвигается нулевая гипотеза о равенстве оцениваемого параметра нулю.

- *коэффициента регрессии в*

H_0: b=0

1. Стандартная ошибка коэффициента регрессии

$$m_b = \sqrt{\frac{\sum (y - \tilde{y})^2}{\sum (x - \bar{x})^2 \cdot (n-2)}}$$

2. Фактическое значение t-критерия Стьюдента

$$t_b = \frac{b}{m_b}$$

3. Определяется $t_{табл}$ (α, $df = n-2$) при определенном уровне значимости и числе степеней свободы.

$t_b > t_{табл}$, H_0 отклоняется и параметр b неслучайно отличается от нуля и сформировался под влиянием систематически действующего фактора *x* (статистически значим с заданной вероятностью).

$t_b < t_{табл}$ – H_0 принимается и признается случайная природа формирования b, т.е. связь надежно не установлена, статистически не значима.

Доверительный интервал для коэффициента регрессии определяется как $b \pm t \cdot m_b$. Поскольку коэффициент регрессии имеет четкую экономическую интерпретацию, доверительные границы интервала не должны содержать противоречивой информации, в частности включать значение 0 в этот интервал.

4. Интерпретация уравнения регрессии

Для прогнозирования возможных значений результативного признака

- авторегрессионное прогнозирование по тренду и колеблемости;

- факторное прогнозирование, основанное на изучении и количественном измерении взаимосвязи между признаками.

Основным условием прогнозирования на основании регрессионного уравнения является стабильность или, по крайней мере, малая изменчивость других факторов и условий изучаемого процесса, не связанных с ними. Если резко изменится «внешняя среда» протекающего процесса, прежнее уравнение регрессии потеряет свое значение.

Прогнозирование по уравнению регрессии проводится в два этапа:

1. Вычисляется «точечный прогноз».
2. Определяется доверительный интервал с достаточно большой вероятностью (интервальный прогноз).

5. Средняя ошибка аппроксимации

Важнейшей характеристикой качества модели является средняя относительная ошибка аппроксимации:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{\hat{y}_i - y_i}{y_i}\right| \cdot 100$$

$\hat{y}_i$, y_i - соответственно расчетное и фактическое значения;

n – число наблюдений.

$\overline{A}$ < 10-13% - высокая точность модели;

10% < $\overline{A}$ < 20% - хорошая точность модели;

20% < $\overline{A}$ < 50% - удовлетворительная.

6. Виды нелинейных зависимостей, поддающиеся линеаризации

Рассмотрим примеры линеаризующих преобразований:

1) Полиномиальная модель: $\hat{y} = a + b_1 x + b_2 x^2 + \cdots + b_p x^p$.

Соответствующая линейная модель: $\hat{y} = a + b_1 z_1 + b_2 z_2 + \cdots + b_p z_p$, где $z_1 = x, z_2 = x^2, \ldots, z_p = x^p$.

2) Гиперболическая модель: $\hat{y} = a + \frac{b}{x}$.

Соответствующая линейная модель: $\hat{y} = a + bz$, где $z = \frac{1}{x}$.

3) Логарифмическая модель: $\hat{y} = a + b \cdot \ln x$.

Соответствующая линейная модель: $\hat{y} = a + bz$, где $z = \ln x$.

Следует отметить и недостаток такой замены переменных, связанный с тем, что вектор оценок получается не из условия минимизации суммы квадратов отклонений для исходных переменных, а из условия минимизации суммы квадратов отклонений для преобразованных переменных, что не одно и то же.

Примеры внутренне линейных моделей и их линеаризация:

1) Мультипликативная степенная модель: $\hat{y} = a x_1^{b_1} x_2^{b_2} \ldots x_p^{b_p}$.

Линеаризующее преобразование:

$\ln \hat{y} = \ln a + b_1 \ln x_1 + b_2 \ln x_2 + \cdots + b_p \ln x_p$

или

$\hat{Y} = A + b_1 z_1 + b_2 z_2 + \cdots + b_p z_p$,

где $\hat{Y} = \ln \hat{y}$, $A = \ln a$, $z_1 = \ln x_1$, $z_2 = \ln x_2, \ldots,$ $z_p = \ln x_p$.

2) Экспоненциальная модель: $\hat{y} = e^{a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}$.

Линеаризующее преобразование: $\ln \hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$.

3) Обратная регрессионная модель: $\hat{y} = \frac{k}{a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}$.

Линеаризующее преобразование: $\frac{1}{\hat{y}} = \frac{a}{k} + \frac{b_1}{k} x_1 + \frac{b_2}{k} x_2 + \cdots + \frac{b_p}{k} x_p$.

К моделям, полученным после проведения линеаризующих преобразований можно применять обычные методы исследования линейной регрессии. Но поскольку в них присутствуют не фактические значения изучаемого показателя, то оценки параметров

получаются несколько смещенными. При анализе линеаризуемых функций регрессии, следует особенно тщательно проверять выполнение предпосылок метода наименьших квадратов.

1.2 Лекция №2 (1 час)

Тема: Множественная регрессия и корреляция

1.2.1 Вопросы лекции:

1. Предпосылки МНК.
2. Множественная линейная регрессия и оценка ее параметров.
3. Проверка статистической значимости уравнения регрессии в целом и его параметров.
4. Показатели корреляции и оценка их значимости.
5. Мультиколлинеарность.

1.2.2. Краткое содержание вопросов

1. Предпосылки МНК

1. Математическое ожидание случайного отклонения ε_i равно нулю: $M(\varepsilon_i) = 0$, для всех наблюдений.

2. Дисперсия случайных отклонений ε_i постоянна: $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ для любого наблюдения $i = j$.

3. Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$.

Предполагается, что отсутствует систематическая связь между любыми случайными отклонениями: $\sigma_{\varepsilon_i\varepsilon_j} = \mathrm{cov}(\varepsilon_i\varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases}$. Если данное условие выполняется, то говорят об отсутствии *автокорреляции*.

4. Случайное отклонение должно быть независимо от объясняющих переменных.

Обычно это условие выполняется автоматически, если объясняющие переменные не являются случайными в данной модели: $\sigma_{\varepsilon_i x_i} = 0$. Выполнимость данной предпосылки не столь критична для эконометрической модели.

5. Модель является линейной относительно параметров.

Для случая множественной регрессии существенными являются еще две предпосылки.

6. Отсутствие мультиколлинеарности.

Между объясняющими переменными отсутствует строгая (сильная) линейная зависимость.

7. Ошибки ε_i , $i=1,2,\dots, n$, имеют нормальное распределение ($\varepsilon_i \sim N(0, \sigma)$).

Выполнимость данной предпосылки важна для проверки статистических гипотез и построения интервальных оценок.

Теорема (Гаусса-Маркова) Если предпосылки 1-7 выполняются, то оценки полученные по МНК обладают свойствами несмещенности, состоятельности и эффективности.

(Бородич С.А. Эконометрика / Учебное пособие, стр. 124)

2. Множественная линейная регрессия и оценка ее параметров

Как и в парной зависимости, возможны разные виды уравнений множественной регрессии: линейные и нелинейные.

Ввиду четкой интерпретации параметров наиболее широко используются линейная функция: $\tilde{y}_x = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$. (*)

В линейной множественной регрессии параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с увеличением соответствующего фактора на единицу при неизменном значении других факторов, закрепленных на среднем уровне.

Основная задача регрессионного анализа заключается в нахождении по выборке объемом n, оценки неизвестных коэффициентов регрессии $b_0, b_1, \ldots, b_k$ модели (*).

Для оценки параметров уравнения множественной регрессии применяют метод наименьших квадратов. Для линейных уравнений строиться система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{cases} \sum y = na + b_1 \sum x_1 + b_2 \sum x_2 + \ldots + b_n \sum x_n \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_2 x_1 + \ldots + b_n \sum x_n x_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum yx_n = a \sum x_n + b_1 \sum x_1 x_n + b_2 \sum x_2 x_n + \ldots + b_n \sum x_n^2 \end{cases}$$

Для ее решения может быть применен метод определителей:

$$a = \frac{\Delta a}{\Delta}, \quad b_1 = \frac{\Delta b_1}{\Delta}, \quad \ldots, \quad b_n = \frac{\Delta b_n}{\Delta},$$

где Δa, $\Delta b_1, \ldots$, Δb_p – частные определители, которые получаются путем замены соответствующего столбца матрицы определителя системы данными левой части системы.

Δ - определитель системы.

Возможна также и другая запись уравнения (*), в так называемом ***стандартизованном масштабе***: $t_Y = \beta_1 t_{x_1} + \ldots + \beta_k t_{x_k} + \varepsilon$,

где $t_y, t_{x_1}, \ldots, t_{x_k}$ - стандартизованные переменные: $t_y = \frac{y - \bar{y}}{\sigma_y}$, $t_{x_j} = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}$, *j=1, 2, ..., k*, для которых среднее значение равно нулю, а среднее квадратическое отклонение равно единице;

β_j – стандартизованные коэффициенты регрессии.

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}, j=1, 2, \ldots, k.$$

Данное соотношение позволяет переходить от уравнения (*) к уравнению в стандартизованном масштабе.

Стандартизованные коэффициенты регрессии показывают на сколько «сигм» изменится в среднем результат (У), если соответствующий фактор X_j изменится на одну «сигму» при неизменном среднем уровне других факторов.

В силу того, что все переменные центрированы и нормированы, коэффициенты β_j *j=l,2,...,k*, сравнимы между собой (в этом их отличие от b). Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат, что позволяет произвести отсев факторов — исключить из модели факторы с наименьшими значениями β_j.

3. Проверка статистической значимости уравнения регрессии в целом и его параметров.

Значимость уравнения регрессии, т. е. гипотеза H_0: b=0 ($b_0 = b_1 = \ldots = b_k = 0$), проверяется по *F-критерию*, наблюдаемое значение которого определяется по формуле:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m}{m-1}$$

R^2 - коэффициент (индекс) множественной детерминации;

m - число параметров;

n - число наблюдений.

По таблице F-распределения для заданных α, V_1= к + 1, V_2 = n – k – 1 находят $F_{кр}$. Гипотеза H_0 отклоняется с вероятностью α, если $F_{набл} > F_{кр}$. Из этого следует, что уравнение является значимым, т. е. хотя бы один из коэффициентов регрессии отличен от нуля.

Для проверки значимости отдельных коэффициентов регрессии, т. е. гипотез H_0: $b_j = 0$, где *j=l,2,...k*, используют t-критерий и вычисляют:

$t_{набл}(b_j) = b_j / m_{bj}$.

$$m_{b_j} = \frac{\sigma_y \cdot \sqrt{1-R_y^2}}{\sigma_{x_j} \cdot \sqrt{1-R_{x_j}^2}} \cdot \frac{1}{\sqrt{n-m-1}},$$

где R_y – коэффициент детерминации уравнения множественной регрессии;

R_{xj} – коэффициент детерминации для зависимости фактора x_j со всеми другими факторами уравнения регрессии.

По таблице t - распределения для заданного α и v = n – k – 1, находят $t_{кр}$.

Гипотеза H_0 отвергается с вероятностью α, если $t_{набл} > t_{кр}$

Из этого следует, что соответствующий коэффициент регрессии b_j значим, т. е. $b_j \neq 0$. В противном случае коэффициент регрессии незначим и соответствующая переменная в модель не включается.

4. Показатели корреляции и оценка их значимости.

Прежде чем определять вид зависимости и описывать ее уравнением регрессии необходимо ответить на вопрос о наличии или отсутствии зависимостей между анализируемыми показателями.

Если распределение значений каждого факторного признака $X_j = (x_1, x_2, \ldots, x_k)_j$ подчиняется *k*-мерному нормальному закону распределения, то степень линейной зависимости между признаками величинами с помощью парных, частных и множественных коэффициентов корреляции и детерминации.

Парный коэффициент корреляции характеризует тесноту линейной зависимости между двумя переменными на фоне действия всех остальных показателей, входящих в модель.

$$r_{jl} = \frac{1/n \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)}{\sigma_j \cdot \sigma_l}, \quad -1 \le r_{jl} \le 1.$$

Если r_{jl} близок к ±1, то связь между переменными сильная, если r_{jl}=0 – линейная связь отсутствует.

$r_{jl} > 0$ – связь между переменными прямая;

$r_{jl} < 0$ – связь обратная.

Для многомерной корреляционной модели важную роль играют частные и множественные коэффициенты корреляции, детерминации (квадраты соответствующих коэффициентов корреляции).

Частный коэффициент корреляции характеризует тесноту линейной связи между двумя переменными при исключении влияния всех остальных признаков, входящих в модель.

Обладает всеми свойствами парного коэффициента корреляции.

Частный коэффициент корреляции k-2-го порядка между признаками x_1 и x_2 при фиксированном воздействии переменных x_3, x_4, …, x_k может быть определен по следующей формуле:

$$r_{x_1x_2}(x_3, x_4, ..., x_k) = r_{12/3,4,...,k} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}}$$

R_{jl} — алгебраическое дополнение матрицы R к элементу r_{jl}.

$$r_{yx_j}(x_1, ..., x_{j-1}, x_{j+1}, ..., x_k) = r_{yx_j/1,...,j-1,j+1,...,k} = \sqrt{1 - \frac{1 - R^2_{yx_1...x_k}}{R^2_{yx_1...x_{j-1},x_{j+1}...x_k}}},$$

$R^2_{yx_1...x_k}$ - множественный коэффициент детерминации всего комплекса *k* факторов с результатом;

$R^2_{yx_1...x_{j-1},x_{j+1},...x_k}$ - тот же показатель детерминации, но без введения в модель фактора x_j.

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. В практических исследованиях предпочтение отдают показателям частной корреляции самого высокого порядка.

Частный коэффициент корреляции между x_1 и x_2 при фиксированном воздействии переменной x_3 рассчитывается по формуле:

$$r_{x_1x_2}(x_3) = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}}$$

Остальные частные коэффициенты определяются аналогично, путем замены соответствующих индексов.

Если частный коэффициент корреляции меньше парного, т.е. $r_{x_1x_2}(x_3) < r_{x_1x_2}$, то взаимодействие между x_1 и x_2 обусловлено частично (или полностью, если $r_{x_1x_2}(x_3) = 0$) воздействием фиксируемых прочих переменных, т.е. - x_3. Если частный коэффициент корреляции $r_{x_1x_2}(x_3) > r_{x_1x_2}$, то фиксируемые прочие переменные ослабляют линейную связь.

<u>*Множественный коэффициент корреляции*</u> характеризует степень линейной связи между одной переменной (результативной) и остальными, входящими в модель, и может быть рассчитан по формуле:

$$r_y = r_y(x_1, x_2, ..., x_k) = \sqrt{1 - \frac{\det R}{R_{yy}}}$$

$$r_y = r_y(x_1, x_2) = \sqrt{1 - \frac{\det R}{R_{yy}}} = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{x_1x_2}r_{yx_1}r_{yx_2}}{1 - r^2_{x_1x_2}}}$$

Если $r_y = 1$, то связь между y и двумерной переменной (x_1, x_2) является функциональной, линейной. Если $r_{x_3} = 0$, то; линейной связи нет.

Из формулы r_y следует, что $r_y > |r_{yx1}|$, $r_y > |r_{yx2}|$, $r_y > |r_{yx1|x2}|$, $r_y > |r_{yx2|x1}|$.

Отсюда можно заметить, что коэффициент множественной корреляции может только увеличиваться, если в модель включать дополнительные признаки – случайные величины, и не увеличиться, если из имеющихся признаков производить исключение.

Если $r_3 = 0$, то $r_{31} = r_{32} = r_{31|2} = r_{32|1} = 0$.

Наибольшему множественному коэффициенту детерминации соответствуют большие частные коэффициенты детерминации (например, r_1^2 соответствуют $r^2_{12|3}$ и $r^2_{13|2}$)

<u>*Множественный коэффициент детерминации*</u> (квадрат соответствующего множественного коэффициента корреляции) характеризует долю дисперсии, например,

случайной величины y, обусловленную изменением остальных случайных величин x_1 и x_2, входящих в модель.

Множественный коэффициент детерминации характеризует качество регрессионной модели.

Однако использование R^2 в случае множественной регрессии является *не вполне корректным,* так как коэффициент детерминации возрастает при добавлении регрессоров в модель. Это происходит потому, что остаточная дисперсия уменьшается при введении дополнительных переменных. И если число факторов приблизится к числу наблюдений, то остаточная дисперсия будет равна нулю, и коэффициент множественной корреляции, а значит и коэффициент детерминации, приблизятся к единице, хотя в действительности связь между факторами и результатом и объясняющая способность уравнения регрессии могут быть значительно ниже.

Для того чтобы получить адекватную оценку того, насколько хорошо вариация результирующего признака объясняется вариацией нескольких факторных признаков, применяют ск*орректированный коэффициент детерминации*

$$R^2_{скор} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}.$$

Скорректированный коэффициент детерминации всегда меньше R^2. Кроме того, в отличие от R^2, который всегда положителен, $R_c{}^2_{корр}$ может принимать и отрицательное значение. Устраняет эффект, связанный с ростом R^2 при возрастании числа регрессоров, являясь коррекцией R^2 на число регрессоров. Использование $R^2_{скор}$ более корректно для сравнения регрессий при изменении количества регрессоров.

Полученные коэффициенты корреляции определяются по выборочной совокупности и являются точечными оценками. Поэтому необходима статистическая проверка значимости коэффициентов корреляции.

Проверка статистической значимости парного и частного коэффициентов корреляции.

Назовем параметр связи в генеральной совокупности значимо отличающимся от нуля (значимым), если гипотеза о равенстве нулю этого параметра отвергается с заданным уровнем значимости α. Если же эта гипотеза принимается, генеральный параметр связи называется незначимым. В корреляционной модели соответствующая связь между величинами считается недоказанной или отсутствующей.

Проверка статистической значимости парного и частного коэффициентов корреляции осуществляется с помощью *t – критерия Стьюдента.*

H_0: ρ=0, H_1: ρ≠0

$$t_{расч} = \frac{r}{\sqrt{1-r^2}} \sqrt{n-l-2} \quad ,$$

где *r* – частный или парный коэффициент корреляции;

l – порядок частного коэффициента корреляции;

Если $|t_{расч}| > t_{табл}(\alpha, v=n-l-2)$, то гипотеза H_0 с вероятностью ошибки 5% отвергается, проверяемый коэффициент корреляции считается значимым с вероятностью 95%.

Если же $|t_{расч}| < t_{табл}(\alpha, v=n-l-2)$, то гипотеза H_0 не отвергается.

С помощью таблиц Фишера-Иейтса.

По таблице Фишера–Иейтса определяется $r_{кр}(\alpha, v=n-l-2)$.

Если $|r_{расч}| > r_{кр}(\alpha, v=n-l-2)$, то гипотеза H_0 с вероятностью ошибки 5% отвергается, проверяемый коэффициент корреляции считается значимым с вероятностью 95%.

Если же $|r_{расч}| < r_{кр}(\alpha, v=n-l-2)$, то гипотеза H_0 не отвергается.

Проверка статистической значимости множественного коэффициента корреляции и коэффициента детерминации проводится с помощью *F-критерия Фишера*.

Нулевая гипотеза: отсутствует линейная связь между переменной x: и остальными переменными, образующими многомерный признак, H_0: $R_y=0$.

Рассчитываем статистику

$$F = \frac{R_y^2 / k}{(1-R_y^2)/(n-k-1)}$$

Если $F_{расч} > F_{табл}$ (α, $v_1=k$, $v_2=n-k-1$), то гипотеза H_0 отвергается, т.е. r_y значимо отличается от нуля, следовательно, линейная связь есть и она статистически значима с вероятностью $\gamma=1-\alpha$.

Иначе связь между случайной величиной *y* и остальными случайными величинами отсутствует.

Конечно, проверку значимости коэффициентов связи начинать с частных коэффициентов корреляции не обязательно. Можно в некоторых случаях сократить такую проверку, например, если r_1 незначим, то коэффициенты $r_{12|3}$ и $r_{13|2}$ становятся незначимыми. Далее, если $r_{12|3}$ незначим, то $r_1=|r_{13}|$ (множественный коэффициент корреляции незначимо отличается от абсолютной величины парного коэффициента корреляции).

Для значимых параметров корреляции можно найти интервальные оценки.

5. Мультиколлинеарность.

Одним из вопросов спецификации модели является отбор факторов, включаемых в модель.

Включение в уравнение множественной регрессии того или иного набора факторов связано, прежде всего, с представлениями исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями.

Факторы, включаемые во множественную регрессию должны отвечать следующим требованиям:

1. Они должны быть количественно измеримы.

2. Отсутствие коллинеарности и мультиколлинеарности факторов.

Одним из основных препятствий эффективного применения множественного регрессионного анализа, является мультиколлинеарность.

Мультиколлинеарность - линейная взаимосвязь двух или нескольких объясняющих переменных.

Последствия мультиколлинеарности:

1. система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии. (Матрица (X^TX) становится слабообусловленной, т.е. ее определитель близок к нулю).

2. большие дисперсии $\widehat{S}_{b_j}^2$ (стандартные ошибки) оценок коэффициентов регрессии. Это расширяет интервальные оценки, ухудшая их точность;

3. заниженные значения $t(b_j)$, что может привести к неоправданному выводу о существенности влияния соответствующего фактора на результат;

4. затрудняется определение влияния каждой из объясняющих переменных на результативный показатель, и параметры параметры уравнения регрессии оказываются неинтерпретируемыми;

5. возможно получение неверного знака у коэффициента регрессии;

6. завышенные значения множественного коэффициента корреляции

На практике, о наличии мультиколлинеарности, обычно судят по матрице парных эффициентов корреляции. Если один из элементов матрицы R больше 0,7 - 0,8, т. е. r_{jl}

>0,7-0,8 , то считают что имеет место мультиколлинеарность. А также о ее проявлениях, указанных в «последствиях мультиколлинеарности».

Устранять мультиколлинеарность или нет – зависит от цели исследования. Если целью является прогноз значений зависимых переменных, то наличие мультиколлинеарности не сказывается на прогнозных значениях, при условии, что между коррелированными переменными будут сохраняться те же отношения. Если цель – определение влияния каждой из объясняющей переменной на зависимую переменную, то мультиколлинеарность – серьезная проблема.

Чтобы избавиться от этого негативного явления при построении эконометрических моделей:

1. используют метод исключения переменных: из уравнения регрессии исключают один или несколько факторных признаков. Предпочтение отдается тому фактору, который при достаточно тесной связи с Y имеет наименьшую тесноту связи с другими факторами. Исключение факторов может проводиться и по t-критерию Стьюдента. Из уравнения исключаются факторы с величиной t-критерия меньше табличного.

Однако необходимо учитывать, что если переменные были включены в модель на теоретической основе, то неправомочно исключать их только ради «улучшения» статистических результатов.

2. используют алгоритм пошагового включения переменных: первоначально строится парное уравнение регрессии с фактором, наиболее тесно связанным с результатом, проверяется значимость уравнения в целом и его параметров. Затем вводится следующая переменная, по силе связи с зависимой переменной. Если вновь вводимая переменная улучшает модель, т.е. значение множественного коэффициента детерминации увеличивается, то ее включают в модель.

3. Сроят уравнение регрессии на главных компонентах.

1.3 Лекция №3 (1 час)

Тема: Моделирование нестационарных временных рядов

1.3.1. Вопросы лекции:

1. Основные элементы временного ряда
2. Модели нестационарных временных рядов и их идентификация
3. Модель авторегрессии – проинтегрированного скользящего среднего
4. Модели рядов, содержащих сезонную компоненту

1.3.2.Краткое содержание вопросов

1. Основные элементы временного ряда

ВР (ряд динамики, динамический ряд) – это последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления.

ВР состоят из двух элементов:

1. периода времени, за который или по состоянию на который приводятся числовые значения (t);
2. числовых значений того или иного показателя, называемых уровнями ряда (y).

В практике исследования динамики явлений и прогнозирования принято считать, что значения уровней временных рядов могут содержать следующие компоненты (структурообразующие компоненты):

- Тренд (u_t);
- Сезонную компоненту (S_t);
- Циклическую компоненту (V_t);
- Случайную компоненту (ε_t).

2. Модели нестационарных временных рядов и их идентификация

Общий вид моделей: $Y = T + S + E$, $Y = T \cdot S \cdot E$.

Выбор одной из 2-х моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно const, строят аддитивную модель, в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда.

Процесс построения модели включает в себя следующие шаги:

1. Выравнивание исходного ряда методом скользящей средней.
2. Расчет значений сезонной компоненты *S*.
3. Устранение сезонной компоненты из исходных уравнений ряда и получение выровненных данных (*T+E)* в аддитивной или (*T·E*) в мультипликативной модели.
4. Аналитическое выравнивание уровней (*T+E)* или (*T·E*) и расчет значений *T* с использованием полученного уравнения тренда.
5. Расчет полученных по модели значений (*T+E)* или (*T·E*).
6. Расчет абсолютных и/или относительных ошибок.

Если полученные значения ошибок не содержат автокорреляции, или можно заметить исходные уровни ряда и в дальнейшем использовать временной ряд ошибок *E* для анализа взаимосвязи исходного ряда и других временных рядов.

3. Модель авторегрессии – проинтегрированного скользящего среднего

В авторегрессии каждое значение ряда находится в линейной зависимости от предыдущих значений. Если анализируемый динамический процесс зависит от значений, отстоящих на *p* временных лагов назад, то авторегрессионный процесс порядка *p*, т.е. AR (*p*):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + \varepsilon_t ,$$

где ε_t - «белый шум» с $\mu_\varepsilon = 0$.

α_0 - свободный член (часто приравнивается к нулю (опускается)).

Для выполнения условия стационарности все корни характеристического уравнения $1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p = 0$ должны быть по модулю больше 1 и различны, т.е. $|z| > 1$. Если $|z| = 1$, процесс называется процессом *единичного корня* и является нестационарным.

Рассмотрим простейший вариант линейного авторегрессионного процесса – *модель авторегрессии 1-го порядка – AR(1), или марковский процесс.*

Эта модель может быть представлена в виде:

$$y_t = \alpha y_{t-1} + \varepsilon_t ,$$

где α - числовой коэффициент, $|\alpha| < 1$,

ε_t - последовательность случайных величин, образующих белый шум.

4. Модели рядов, содержащих сезонную компоненту

Выравнивание исходного ряда методом скользящей средней:

1) Для этого суммируется, уровни ряда последовательно за каждые 4 квартала со сдвигом на 1 момент времени и определяются условные годовые объемы, уровни и т.д.

2) Разделив полученные суммы на 4, найдем скользящие средние. Полученные т.о. выровненные значения уже не содержат сезонной компоненты.

3) Приведем эти значения в соответствии с фактическими моментами времени, для чего найдем средние значения из двух последовательных скользящих средних – центрированные скользящие средние.

Оценки сезонной компоненты найдем как разность между фактическими уровнями ряда и центрированными скользящими средними. Используем эти оценки для расчета значений сезонной компоненты *S*. Для этого найдем средние за каждый квартал оценки сезонной компоненты S_i. В моделях с сезонной компонентой обычно предполагается, что сезонные воздействия за период взаимопогашаются. Например, в аддитивной модели это выражается в том, что сумма значений сезонной компоненты по всем кварталам должна быть равна нулю.

В мультипликативной модели взаимопогашаемость сезонных воздействий выражается в том, что сумма значений сезонной компоненты по всем кварталам должна быть равна числу периодов в цикле. Например, при сезонных колебаниях число периодов одного цикла (год) равно 4 (4 квартала).

В аддитивной модели: $S_i = \overline{S}_i - K$, где *K* – корректирующий коэффициент, $K = \sum_{i=1}^{4} S_i / 4$ и тогда $\sum S_i = 0$;

В мультипликативной модели: $S_i = \overline{S}_i \cdot K, i = \overline{1,4}$, $K = 4 \Big/ \sum_{i=1}^{4} S_i$ и тогда $\sum_{i=1}^{4} S_i = 4.$

В аддитивной модели вычитаем ее значения из каждого уровня исходного временного ряда. Получим $T+E=Y-S$.

В мультипликативной модели разделим каждый уровень исходного ряда на соответствующие значения сезонной компоненты. Тем самым получим $T \cdot E = Y \div S$, которые содержат только тенденцию и случайную компоненту.

Определение компоненты *T* данной модели.

Для этого проводится аналитическое выравнивание ряда $T+E$ $(T \cdot E)$ с помощью линейного тренда $T = a + bt.$

Для этого посчитаем *a* – const, *b* – коэффициент регрессии, стандартную ошибку коэффициента регрессии R^2, число наблюдений и число степеней свободы. С помощью них определяем значимость регрессии.

Найдем значения уровней ряда по T_i – полученным по теоретической (аналитической) формуле и S_i – значениям сезонной компоненты для соответствующих кварталов.

$$E = Y - (T + S)$$

$$E = Y \div (T \cdot S)$$

Это абсолютные значения (абсолютные ошибки). По аналогии с моделью регрессии для оценки качества построенной модели или для выбора наилучшей модели можно применять сумму квадратов полученных абсолютных ошибок.

$\left(1 - \frac{\sum E_i^2}{\sum (y_i - \bar{y})^2}\right) \cdot 100\%$ – это доля факторной дисперсии уровней ряда объясняет полученное количество процентов от общей вариации уровней временного ряда.

1.4 Лекция №4 (1 час)

Тема: Анализ временных рядов

1.4.1. Вопросы лекции:

1. Стационарные временные ряды и их основные характеристики
2. Модели стационарных временных рядов

1.4.2.Краткое содержание вопросов

1. Стационарные временные ряды и их основные характеристики

Стохастический процесс Y_t называется *стационарным в сильном смысле (строго стационарным или стационарным в узком смысле)*, если совместное распределение вероятностей всех переменных $y_{t1}, y_{t2},..., y_{tn}$ точно то же самое, что и для переменных $y_{t1+\tau}, y_{t2+\tau},..., y_{tn+\tau}$.

Под стационарным процессом *в слабом смысле* (в широком смысле) понимается стохастический процесс, для которого среднее и дисперсия независимо от рассматриваемого периода времени имеют постоянное значение, а автокорреляция зависит только от длины лага между рассматриваемыми переменными:

$$\mu(y_t) = \mu(y_{t+\tau}) = \mu;$$
$$D(y_t) = D(y_{t+\tau}) = \mu(y_t - \mu)^2 = \mu(y_{t+\tau} - \mu)^2 = D_0 = const;$$
$$\operatorname{cov}(y_t, y_{t+\tau}) = \mu[(y_t - \mu)(y_{t+\tau} - \mu)] = \operatorname{cov}(\tau).$$

Из этого следует, что автоковариация будет зависеть только от сдвига по времени τ и не будет зависеть от t.

При анализе изменения $\operatorname{cov}(\tau)$ в зависимости от временного сдвига τ принято говорить об автоковариационной функции.

С понятием автоковариационной функции тесно связано понятие автокорреляционной функции (АКФ):

$$\rho(\tau) = \frac{\operatorname{cov}(y_t, y_{t+\tau})}{D(y_t)} = \frac{\operatorname{cov}(\tau)}{D_0}$$

Выборочная оценка коэффициента автокорреляции $r(\tau)$ может быть определена следующим образом:

$$r(\tau) = \frac{\frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})},$$

где n - длина временного ряда;

τ - временной сдвиг (лаг);

$\bar{y}$ - среднее значение временного ряда.

Формула для расчета выборочной оценки частного коэффициента автокорреляции:

$$r_{ij.k} = \frac{r_{ij} - r_{ik} \cdot r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}.$$

Например, $r_{ч}$ 1- го порядка между y_t и y_{t+2} при устранении влияния y_{t+1}:

$$r_ч(2) = r_{02.1} = \frac{r(2) - r(1) \cdot r(1,2)}{\sqrt{(1 - r^2(1))(1 - r^2(1,2))}},$$

где $r(2)$ - коэффициент корреляции между y_t и y_{t+2};

$r(1)$- коэффициент корреляции между y_t и y_{t+1};

$r(1,2)$- коэффициент корреляции между y_{t+1} и y_{t+2}.

В практической аналитической работе стационарность временного ряда означает отсутствие:

– тренда;
– систематических изменений дисперсии;
– строго периодичных флуктуаций;
– систематически изменяющихся взаимосвязей между элементами временного ряда.

2. Модели стационарных временных рядов

Наиболее распространенные модели стационарных рядов – *модели авторегрессии и модели скользящего среднего*.

В авторегрессии каждое значение ряда находится в линейной зависимости от предыдущих значений. Если анализируемый динамический процесс зависит от значений, отстоящих на *p* временных лагов назад, то авторегрессионный процесс порядка *p*, т.е. AR (*p*):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + \varepsilon_t,$$

где ε_t - «белый шум» с $\mu_\varepsilon = 0$.

α_0 - свободный член (часто приравнивается к нулю (опускается)).

Для выполнения условия стационарности все корни характеристического уравнения $1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p = 0$ должны быть по модулю больше 1 и различны, т.е. $|z| > 1$. Если $|z| = 1$, процесс называется процессом *единичного корня* и является нестационарным.

Рассмотрим простейший вариант линейного авторегрессионного процесса – *модель авторегрессии 1-го порядка – AR(1), или марковский процесс.*

Эта модель может быть представлена в виде:

$$y_t = \alpha y_{t-1} + \varepsilon_t,$$

где α - числовой коэффициент, $|\alpha| < 1$,

ε_t - последовательность случайных величин, образующих белый шум.

Основные свойства Марковского процесса:

$$\mu y_t = 0$$

$$D(y_t) = \frac{\sigma_0^2}{1 - \alpha^2}$$

$$\operatorname{cov}(y_t, y_{t \pm k}) = \alpha^k D(y_t)$$

$$\rho(y_t, y_{t \pm k}) = \alpha^k$$

Поэтому степень тесноты корреляционной связи между членами последовательности экспоненциально убывает по мере их взаимного удаления друг от друга во времени.

Таким образом, $\alpha = \frac{\operatorname{cov}(y_t, y_{t-1})}{D(y_t)}$

Условие стационарности ряда для AR(1) определяется требованием к коэффициенту $\alpha: |\alpha| < 1$.

Практические рекомендации по идентификации авторегрессионных моделей опираются на изучение АКФ и ЧАКФ:

1. У моделей AR(p) значения АКФ экспоненциально затухают (либо монотонно, либо попеременно меняя знак);

2. ЧАКФ для моделей AR(p) будет иметь ненулевые значения лишь при $k \le p$, а начиная с лага $k = p + 1$ теоретическая ЧАКФ равна нулю. Это свойство становится ключевым при подборе порядка p авторегрессионной модели для конкретных экономических временных рядов.

2. Методические указания по проведению практических занятий

2.1. Практическое занятие 1. Парная регрессия и корреляция (1 час)

2.1.1.Задание для работы:

1. Отбор факторов и оценивание неизвестных параметров классической модели линейной регрессии
2. Фиктивные переменные во множественной регрессии
3. Виды нелинейной регрессии. Оценка параметров
4. Корреляция для нелинейной регрессии

2.1.2 Краткое описание проводимого занятия

Задача 1. По территориям региона приводятся данные за 199X г.:Номер региона Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x Среднедневная заработная плата, руб., y

1	x1	y1
2	x2	y2
3	x3	y3
4	x4	y4
5	x5	y5
6	x6	y6
7	x7	y7
8	x8	y8
9	x9	y9
10	x10	y10
11	x11	y11
12	x12	y12

1. Построить линейное уравнение парной регрессии y от x
2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппрксимации
3. Оценить статистическую значимость параметров регрессии и корреляции.
4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x, составляющем 107% от среднего уровня.
5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Задача 2. По 30 территориям России имеются данные, представленные в таблице.

По данным таблицы:

1. Построить уравнение множественной регрессии в стандартизованной и естественной форме; рассчитать частные коэффициенты эластичности, сравнить их с ß1 и ß2, пояснить различия между ними.

2. Рассчитать линейные коэффициенты частной корреляции и коэффициент множественной корреляции, сравнить их с линейными коэффициентами парной корреляции, пояснить различия между ними.

3. Рассчитать общий и частные F-критерии Фишера.

2.1.3.Результаты и выводы:

Усвоение студентами знаний и навыков по теме практического занятия.

2.2. Практическое занятие 2. Множественная регрессия и корреляция (1 час)

2.2.1.Задание для работы:

1.Отбор факторов при построении множественной регрессии

2. Выбор формы уравнения регрессии

3.Оценка параметров уравнения множественной регрессии

2.2.2 Краткое описание проводимого занятия

1. Рассчитайте параметры линейного уравнения множественной регрессии с полным перечнем факторов по данным в соответствии с вариантом.

2. Дайте сравнительную оценку силы связи факторов с результатом с помощью средних (общих) коэффициентов эластичности.

3. Оцените с помощью F-критерия Фишера-Снедекора значимость уравнения линейной регрессии и показателя тесноты связи. Определите множественный коэффициент корреляции и детерминации и скорректированный коэффициент детерминации.

4. Оцените статистическую значимость коэффициентов регрессии с помощью t-критерия Стьюдента.

5. Оцените качество уравнения через среднюю ошибку аппроксимации.

6. Рассчитайте матрицу парных коэффициентов корреляции и отберите информативные факторы в модели. Укажите коллинеарные факторы.

7. Постройте модель в естественной форме только с информативными факторами и оцените ее параметры.

8. Постройте модель в стандартизованном масштабе и проинтерпретируйте ее параметры.

9. Рассчитайте прогнозное значение результата, если прогнозное значение факторов составляют 80% от их максимальных значений.

10. Рассчитайте ошибки и доверительный интервал прогноза для уровня значимости $\alpha = 0{,}05$.

11. По полученным результатам сделайте экономический вывод.

2.2.3. Результаты и выводы:

Усвоение студентами знаний и навыков по теме практического занятия.

2.3. Практическое занятие 3. Моделирование нестационарных временных рядов (1 час)

2.3.1. Задание для работы:

1. Модели нестационарных временных рядов и их идентификация.

2. Модель авторегрессии – проинтегрированного скользящего среднего; модели рядов, содержащих сезонную компоненту

2.3.2 Краткое описание проводимого занятия

Для временного ряда финансового или социально-экономического показателя с помесячной или поквартальной динамикой Краткое описание проводимого занятия

1) на основе графического анализа провести исследование компонентного состава временного ряда;

2) при обнаружении тенденции во временном ряду оценить параметры линейного и параболического тренда;

3) построить тренд – сезонную аддитивную и мультипликативную модель. Для всех построенных моделей с помощью средней относительной ошибки аппроксимации оценить их качество и дать прогноз на следующие два периода;

4) Оценить автокорреляцию для построенного уравнения регрессии.

Для всех построенных моделей с помощью средней относительной ошибки аппроксимации оценить их качество и дать прогноз на следующие два периода.

2.3.3.Результаты и выводы:

Усвоение студентами знаний и навыков по теме практического занятия.

2.4. Практическое занятие 4. Системы линейных одновременных уравнений (1 час)

2.4.1. Задание для работы:

1. Структурная и приведенная формы модели систем одновременных уравнений.

2. Рекурсивные системы одновременных уравнений. Модель спроса – предложения как пример системы одновременных уравнений.

3. Основные структурные характеристики моделей.

4. Условия идентифицируемости уравнений системы. Идентификация рекурсивных систем.

2.4.2 Краткое описание проводимого занятия

Задача: Идентифицировать следующую систему одновременных уравнений:

$$\begin{cases} \hat{y}_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 \\ \hat{y}_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 \\ \hat{y}_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 \end{cases}$$

2.4.3.Результаты и выводы:

Усвоение студентами знаний и навыков по теме практического занятия.

Разработал(и): _____ Т.В.Тимофеева