

## 1. Линейная модель множественной регрессии. Основные предпосылки метода наименьших квадратов.

$$y_i = \alpha_0 + \alpha_1 \cdot x_{i1} + \alpha_2 \cdot x_{i2} + \dots + \alpha_k \cdot x_{ik} + \varepsilon_i$$
$$i = 1, \dots, n$$

где  $n$  - количество наблюдений,  $k$  - количество факторов, включенных в модель.

Как и в случае парной регрессии, мы так выбираем значения коэффициентов регрессии, чтобы обеспечить наилучшее соответствие наблюдениям в надежде получить оптимальные оценки для неизвестных истинных значений параметров. Как и прежде, оценка оптимальности соответствия определяется минимизацией суммы квадратов отклонений:  $\sum_i (y_i - \hat{y}_i)^2 = \sum e^2 \rightarrow \min$

Для того чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов, давал наилучшие из всех возможных результатов, должны выполняться следующие условия (предпосылки), известные как условия Гаусса – Маркова.

### Основные предпосылки относительно регрессионной модели:

- \* Наблюдений должно быть больше, чем оцениваемых коэффициентов (желательно  $n > 2 \cdot k$ ).
- \* Модель правильно специфицирована
- \* Между эндогенной переменной  $Y$  и  $k$  регрессорами существует линейная регрессионная зависимость.
- \* Регрессионное уравнение линейно по своим параметрам

В матричном это можно записать в следующем виде:  $y = XA + e$ .

С помощью МНК оценивается регрессия  $y$ :

$$A = (X'X)^{-1}X'y = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \dots \\ \hat{\alpha}_k \end{pmatrix}$$

## 2. Нелинейные модели регрессии. Походы к оцениванию. Примеры

Нелинейные модели регрессии используются для аппроксимации сложных зависимостей между переменными в регрессионном анализе. В отличие от линейной регрессии, где зависимость предполагается линейной, нелинейные модели регрессии могут учитывать более сложные нелинейные взаимосвязи. Существует несколько подходов к оцениванию нелинейных моделей регрессии. Рассмотрим некоторые из них: Метод наименьших квадратов (МНК): Это классический метод оценивания параметров модели регрессии, который минимизирует сумму квадратов ошибок между фактическими и прогнозными значениями. Для нелинейных моделей МНК может быть применен с помощью итеративных численных методов, таких как метод Гаусса-Ньютона или метод Левенберга-Марквардта. Метод максимального правдоподобия (ММП): Этот метод основан на максимизации функции правдоподобия, которая измеряет вероятность получения наблюдаемых данных при заданных параметрах модели. Для нелинейных моделей оценка параметров может быть выполнена с использованием итеративных численных алгоритмов, таких как метод Ньютона-Рафсона или метод Бройдена-Флетчера-Гольдфарба-Шанно. Нелинейные обобщенные модели регрессии (НОМР): Это класс методов, который объединяет нелинейные модели с обобщенными линейными моделями (ОЛМ). ОЛМ позволяют моделировать зависимость между предикторами и откликом с помощью нелинейной функции связи, при этом распределение отклика может быть специфицировано с использованием различных семейств вероятностных распределений. НОМР предоставляют более гибкий подход к оцениванию нелинейных моделей регрессии. Пример Предположим, у вас есть файл с данными о продажах автомобилей, включающий информацию о марке, годе выпуска, пробеге и цене автомобиля. Вы хотите построить модель, которая предсказывает цену автомобиля на основе его характеристик. Вы решаете использовать нелинейную модель регрессии для учета нелинейных взаимосвязей между характеристиками автомобиля и его ценой. Например, вы можете предположить, что зависимость между годом выпуска и ценой автомобиля имеет квадратичную форму. 1. Загрузите данные из файла прошлого года в статистическое программное обеспечение или язык программирования, такие как Python или R. 2. Подготовьте данные, удалив выбросы, заполнив пропущенные значения и масштабируя переменные при необходимости. 3. Определите нелинейную модель регрессии, которая наилучшим образом описывает зависимость между предикторами (например, годом выпуска и пробегом) и откликом (ценой автомобиля). Например, модель может иметь следующий вид: 
$$\text{цена} = \beta_0 + \beta_1 * \text{год} + \beta_2 * \text{год}^2 + \beta_3 * \text{пробег},$$
 где  $\beta_0, \beta_1, \beta_2, \beta_3$  - коэффициенты модели, а год и пробег - предикторы. 4. Оцените параметры модели с использованием выбранного метода оценивания, такого как МНК или ММП. Для нелинейной модели регрессии может потребоваться использование итеративных численных алгоритмов. 5. Оцените качество модели, проанализировав результирующие коэффициенты, статистическую значимость предикторов и остатки модели. Выполните диагностику модели для проверки предположений о распределении остатков, гомоскедастичности и отсутствии мультиколлинеарности.

### 3. Тестирование правильности выбора спецификации: типичные ошибки спецификации модели, Тест Рамсея (тест RESET), условия применения теста.

#### Типичные ошибки спецификации модели:

- \* Неверно выбран тип уравнения регрессии
- \* В линейное уравнение множественной регрессии включен несущественный регрессор
- \* В линейное уравнение множественной регрессии не включен существенный регрессор

Все три компонента можно объединить одним словосочетанием: тупые студенты :)

Этот тест обычно применяется в тех случаях, когда экономическая теория или предварительный анализ данных позволяют нам предположить, что зависимость  $Y$  от некоторых переменных является не линейной, а полиномиальной (на практике в подавляющем большинстве случаев — квадратичной). В этом случае пропущенными переменными являются соответствующие степени этих переменных. (источник: Демидова)

Если модель верна, то добавление нелинейных

функций  $\hat{y}_t = x_t' \beta$  не должно помогать объяснять  $y_t$ . В частности, можно добавлять степени:

$$y_t = x_t' \beta + \alpha_2 \hat{y}_t^2 + \alpha_3 \hat{y}_t^3 + \dots + \alpha_m \hat{y}_t^m + \varepsilon_t. \quad (4.28)$$

(источник: Магнус)

#### Тест Рамсея (тест на функциональную форму).

$H_0$ : модель правильно специфицирована

$H_1$ : модель неправильно специфицирована

#### 1. Оценивается спецификация исследуемой модели:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \varepsilon_t, t = 1, \dots, n$$

$$\hat{Y}_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt}$$

#### 2. Оценивается вспомогательная регрессия

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \beta_{p+1} \hat{Y}_t^2 + \dots + \beta_p \hat{Y}_t^{p+1} + v_t$$

$p$  обычно берут равным 2-3, для множественной регрессии - равным 4 (источник: Гринева)

Основная идея этого шага состоит в том, что нелинейные степени  $\hat{Y}$  отражают влияние нелинейных степеней объясняющих переменных (источник: Демидова)

#### 3. Вычисляется статистика:

$$F = \frac{(ESS_1 - ESS_2)/p}{ESS_2/(n - k - p)} \sim F(p, n - k - p)$$

#### 4. Проводится сравнение

$$F > F(p, n - k - p)$$

Если неравенство справедливо, то гипотеза отвергается и спецификация модели признается неверной

F-статистики в пункте 3 имеют распределение Фишера только в случае, если случайные возмущения регрессионной модели являются независимыми и нормально распределенными (источник: Бабешко, Бич, Орлова)

#### 4. Тестирование правильности выбора спецификации: типичные ошибки спецификации модели, Критерий Акаике, Критерий Шварца. Условия применения критериев.

##### Типичные ошибки спецификации модели:

- \* Неверно выбран тип уравнения регрессии
  - \* В линейное уравнение множественной регрессии включен несущественный регрессор
- 

- \* В линейное уравнение множественной регрессии не включен существенный регрессор

Все три компоненты можно объединить одним словосочетанием: тупые студенты :)

##### Критерий Акаике

Рассчитывается статистика Акаике:

$$AIC = \ln\left(\frac{ESS_k}{n}\right) + \frac{2k}{n} + 1 + \ln(2\pi)$$

При увеличении объясняющих переменных первое слагаемое в правой части уменьшается, второе – увеличивается.

Среди нескольких альтернативных моделей (полной и редуцированной) *предпочтение отдается модели, с наименьшим значением статистики AIC*, в которой достигается определенный компромисс между величиной остаточной суммы квадратов и количеством объясняющих переменных.

##### Критерий Шварца.

Рассчитывается статистика Шварца:

$$SC = \ln\left(\frac{ESS_k}{n}\right) + \frac{k \ln(n)}{n} + 1 + \ln(2\pi)$$

При увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а второе – увеличивается.

Среди нескольких альтернативных моделей (полной и редуцированной) *предпочтение отдается модели с наименьшим значением статистики Шварца*.

Данные критерии подходят только для сравнения моделей с одинаковой зависимой переменной.

## 5. Гетероскедастичность: определение, причины, последствия. Тест Голдфелда-Квандта и особенности его применения.

Гетероскедастичность – непостоянство дисперсии отклонений.

Причины:

- Эффект масштаба
- В пространственно-временных данных – эффект запаздывания данных

Последствия:

- Оценки перестают быть эффективными

Увеличение дисперсии оценок снижает вероятность максимально точных оценок.

Поэтому все выводы, полученные на основе  $t$ ,  $F$  статистик и интервальные оценки будут ненадежными

Гомоскедастичность есть постоянство дисперсий всех наблюдений, а так же равенство матожидания этих остатков нулю (условия Гаусса-Маркова)

Для проверки исходных данных на гомоскедастичность существует множество специальных тестов. Одним из них является тест Голдфелда-Квандта.

Тест Голдфелда-Квандта:

Шаг 1.

Все наблюдения упорядочиваются по величине  $x$ .

Шаг 2.

Вся выборка в начале и у конце делится на две части. Количество наблюдений в этих подвыборках одинаково и определяется в соответствии изначальному количеству наблюдений.

Шаг 3.

Получившиеся выборки оцениваются регрессией.

Шаг 4.

По каждой подвыборке рассчитывается сумма квадратов отклонений/

Шаг 5.

Рассчитывается  $F_{\text{расч}}$ -статистика Фишера:

$$F_{\text{расч}} = \frac{S_{\text{бол}}}{S_{\text{мен}}}, \text{ где } S - \text{сумма больших и меньших квадратов отклонений из подвыборок}$$

Шаг 6.

Сравниваются значения  $F_{\text{расч}}$  и  $F_{\text{крит}}(\alpha, \nu_1 = \nu_2 = k - m - 1) \Rightarrow$  гетероскедастичность

## 6. Гетероскедастичность: определение, причины, последствия. Тест ранговой корреляции Спирмена и особенности его применения.

Гетероскедастичность – непостоянство дисперсии отклонений.

Причины:

- Эффект масштаба
- В пространственно-временных данных – эффект запаздывания данных

Последствия:

- Оценки перестают быть эффективными

Увеличение дисперсии оценок снижает вероятность максимально точных оценок.

Поэтому все выводы, полученные на основе  $t$ ,  $F$  статистик и интервальные оценки будут ненадежными

При выполнении теста ранговой корреляции Спирмена предполагается, что дисперсия случайного члена будет либо увеличиваться, либо уменьшаться по мере увеличения  $x$ , и поэтому в регрессии, оцениваемой с помощью МНК, абсолютные величины остатков и значения  $x$  будут коррелированы.

Алгоритм теста:

1. Данные по  $x$  упорядочиваются по возрастанию, каждому значению  $x$  присваивается ранг ( $\text{rang } x$ ).
2. По данным столбца остатков  $e$  строится вспомогательный столбец  $|e|$  (модуль  $e$ ) и также каждому значению этого столбца присваивается ранг ( $\text{rang } |e|$ ).
3. Считается коэффициент ранговой корреляции:

$$r_{xe} = 1 - \frac{6 \sum (D_i^2)}{n(n^2 - 1)},$$
 где  $D_i$  – разность между рангом  $x_i$  и рангом  $e_i$ , т.е.  $D_i = \text{rang } x - \text{rang } |e|$ ,  $n$  – количество наблюдений в выборке.

4. Затем найденный коэффициент ранговой корреляции проверяется на значимость. Для этого вычисляется статистический критерий  $t_{\text{факт}} = \frac{r_{xe}}{\sqrt{1 - r_{xe}^2}} \sqrt{n - k - 1}$  и сравнивается с  $t_{\text{табл}}$ , которое берется из специальной таблицы «распределение Стьюдента» и находится на пересечении чисел  $\gamma = 0.95$  или  $\alpha = 0.05$  и  $(n - k - 1)$ , где  $k$  – число объясняющих переменных в задаче. Если  $t_{\text{факт}} > t_{\text{табл}}$ , то коэффициент корреляции значим, и в исследуемой модели присутствует гетероскедастичность, и наоборот.

## 7. Гетероскедастичность: определение, причины, последствия. Тест Тест Бреуша-Пагана и особенности его применения.

Гетероскедастичность – непостоянство дисперсии отклонений.

Причины:

- Эффект масштаба
- В пространственно-временных данных – эффект запаздывания данных

Последствия:

- Оценки перестают быть эффективными

Увеличение дисперсии оценок снижает вероятность максимально точных оценок.

Поэтому все выводы, полученные на основе t, F статистик и интервальные оценки будут ненадежными

Тест Бреуша-Пагана  $H_0$ : остатки гомоскедастичности  $\gamma_2 = \gamma_p = 0$

1. Коэффициенты регрессии определяются МНК
2. Находим дисперсию ошибки модели  $\hat{\sigma}^2 = \frac{1}{n} RSS$
3. Вычисляем стандартизированные остатки  $\frac{\varepsilon_i^2}{\hat{\sigma}^2}$
4. Строится дополнительная регрессия квадратов стандартизированных ошибок на исходные наблюдаемые переменные:  $\hat{\varepsilon}_i^2 = \gamma_1 + \gamma_2 z_{2i} + \dots + \gamma_p z_{pi} + \eta_i$
5.  $LM = n \cdot R^2$ , где  $R^2$  — коэффициент детерминации построенной на предыдущем шаге регрессии

При справедливости нулевой гипотезы о гомоскедастичности остатков статистика критерия имеет распределение хи-квадрат с p-1 степенями свободы.

## 8. Гетероскедастичность: определение, причины, последствия. Тест Глейзера и особенности его применения.

Гетероскедастичность – непостоянство дисперсии отклонений.

Причины:

- Эффект масштаба
- В пространственно-временных данных – эффект запаздывания данных

Последствия:

- Оценки перестают быть эффективными

Увеличение дисперсии оценок снижает вероятность максимально точных оценок.

Поэтому все выводы, полученные на основе t, F статистик и интервальные оценки будут ненадежными

### Тест Глейзера

1. Строится уравнение регрессии:  $\hat{y}_i = b_0 + \sum_{j=1}^m b_j x_{ij}$  и вычисляются остатки  $e_i = y_i - \hat{y}_i$ ,  $i=1, \dots, n$
2. Выбирается фактор пропорциональности X и оценивают вспомогательное уравнение регрессии:  $|e_i| = \alpha_0 + \alpha_1 * x_i^\gamma + \eta_i$ ,  $i=1, \dots, n$
3. Статистическая значимость коэффициента  $\alpha_1$  в каждом случае означает наличие гетероскедастичности
4. Если для нескольких моделей будет получена значимая оценка  $\alpha_1$ , то характер гетероскедастичности определяют по наиболее значимой из них.

## 9. Способы корректировки гетероскедастичности: взвешенный метод наименьших квадратов (ВМНК) и особенности его применения.

При подтверждении наличия гетероскедастичности необходимо преобразовать модель с целью смягчения ее влияния. Для этого предлагают применять МВНК (метод взвешенных наименьших квадратов).

Этапы МВНК:

- Каждое из наблюдений  $y_i$ ,  $x_i$  делят на известную величину  $\sigma_i$  (дисперсия случайных отклонений известна -  $\sigma_i = \sqrt{\sigma_j^2}$ ; дисперсия случайных отклонений неизвестна -  $\sigma_i = \sqrt{x_i}$  при пропорциональности дисперсии и  $x_i$  или  $\sigma_i = x_i$  при пропорциональности дисперсии и  $x_i^2$ )
- По МНК для преобразования значений строится уравнение регрессии с гарантированным качеством оценок (остатки гомоскедастичны)

## 10. Автокорреляция: определение, причины, последствия. Тест Дарбина-Уотсона и особенности его применения.

Автокорреляция - это наличие сильной корреляционной зависимости между последовательными наблюдениями одного ряда.

Виды автокорреляции:

- положительная (на графике проявляется чередованием зон положительных и отрицательных остатков).
- отрицательная (остатки "слишком часто" меняются).

Основными причинами автокорреляции являются:

- Спецификации (неправильный выбор формы модели).
- Инерция.
- Эффект "Паутины" (реакция экономических показателей на изменение условий с запазданием).

Последствия автокорреляции:

- Автокорреляция не приводит к смещению оценок регрессии, но оценки перестают быть эффективными.
- Автокорреляция (особенно положительная) часто приводит к уменьшению стандартных ошибок коэффициентов, что влечет за собой увеличение t-статистик
- Оценка дисперсии остатков является смещенной оценкой истинного значения.

В силу вышесказанного выводы по оценке качества коэффициентов и модели в целом, возможно, будут неверными. Это приводит к ухудшению прогнозных качеств модели.

Для определения автокорреляции в рассматриваемой модели можно использовать критерий Дарбина-Уотсона.

Ограничения Дарбина-Уотсона:

- Модель должна содержать свободный член.
- Данные должны иметь одинаковую периодичность.
- Дарбин-Уотсон не применим к моделям с автокорреляцией.

Статистика Дарбина-Уотсона:

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

Значение d-статистики сравнивается с критическими значениями  $d_1$  и  $d_2$ . При этом могут возникнуть следующие ситуации:

- если  $d_2 < d < (4 - d_2)$ , то остатки признаются некоррелированными;
- если  $0 < d < d_1$ , то имеется положительная автокорреляция;
- если  $(4 - d_1) < d < 4$ , то существует отрицательная автокорреляция;
- если  $d_1 < d < d_2$  или  $(4 - d_2) < d < (4 - d_1)$ , то это указывает на неопределенность ситуации.

## 11. Автокорреляция: определение, причины, последствия. Тест Бройша – Годфри и особенности его применения.

Зависимость случайных возмущений от времени называется *автокорреляцией*. Автоковариационная матрица вектора случайных возмущений при наличии автокорреляции имеет структуру:

$$\Omega = \begin{pmatrix} \sigma^2 & \Omega_{12} & \dots & \Omega_{1n} \\ \Omega_{21} & \sigma^2 & \dots & \vdots \\ \vdots & \vdots & \sigma^2 & \vdots \\ \Omega_{n1} & \dots & \dots & \sigma^2 \end{pmatrix} \quad \Omega_{t,s} = \text{Cov}(\varepsilon_t, \varepsilon_s) \neq 0$$

Причины: ошибки спецификации модели (пропуск важной объясняющей переменной, использование ошибочной функциональной зависимости между переменными); ошибки измерений; характер наблюдений (например, данные временных рядов: если существует корреляция между

последовательными значениями некоторой независимой переменной, то будет наблюдаться и корреляция последовательных значений остатков).

Последствия: Оценки параметров, оставаясь линейными и несмещенными, перестают быть эффективными.

Следовательно, они перестают обладать свойствами наилучших линейных несмещенных оценок.

Стандартная оценка дисперсии случайных ошибок смещена вниз в случае положительной автокоррелированности ошибок и смещена вверх, если автокорреляция ошибок отрицательна.

Оценка дисперсии регрессии является смещенной оценкой истинного значения, во многих случаях занижая его. В силу вышесказанного выводы по t- и F-статистикам, определяющим значимость коэффициентов регрессии и коэффициента детерминации, возможно, будут неверными. Вследствие этого ухудшаются прогнозные качества модели.

Построенные доверительные интервалы для стандартных отклонений оценок коэффициентов не соответствуют заявленным уровням значимости: в случае положительной автокорреляции построенные интервалы неоправданно узки, в противном случае – широки.

### Тест Бреуша-Годфри

- \* требуется большой объем выборки
- \* позволяет выявить автокорреляцию для случая  $p \geq 1$
- \* допускает включение лаговых переменных
- \* вообще говоря, требует, чтобы остатки были гомоскедастичны, но на практике на это закрывают глаза

Тест является асимптотическим, т.е. для достоверности выводов требуется большой объем выборки. В тесте рассматривается авторегрессия остатков на их лаговые значения. Пусть спецификация модели множественной регрессии имеет вид:

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \varepsilon_t, \quad t = 1, \dots, n, \quad (3.3)$$

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + \rho_p \varepsilon_{t-p} + v_t,$$

где  $v_t$ ,  $t = 1, \dots, n$  — независимые в совокупности и от  $\varepsilon_{t-j}$ ,  $j = 1, \dots, p$ , случайные возмущения, имеющие одинаковое нормальное распределение  $N(0, \sigma_v^2)$ . Статистика теста формируется по правилу

$$BG = nR^2, \quad (3.4)$$

где  $R^2$  — коэффициент детерминации вспомогательной авторегрессионной модели

$$e_t = b_1 X_{1t} + b_2 X_{2t} + \dots + b_k X_{kt} + r_1 e_{t-1} + r_2 e_{t-2} + \dots + r_p e_{t-p} + u_t, \quad (3.5)$$

$e_t$  — остатки, полученные в результате МНК оценивания основной регрессионной модели (3.3),  $e_{t-j} = 0$  при  $t < j + 1$ . Если нулевая гипотеза

$$H_0: r_1 = r_2 = \dots = r_p = 0$$

верна, то при большом значении  $n$  статистика (3.4) имеет хи-квадрат распределение с параметром  $p$ , и гипотеза не отклоняется при уровне значимости  $\alpha$ , если выполняется неравенство

$$BG < \chi_{\alpha}^2(p).$$

Для авторегрессии первого порядка

$$e_t = 0,057 - 0,033 \cdot X_t + 0,456 \cdot e_{t-1} + \hat{u}_t, \quad \begin{matrix} (0,173) & (0,105) & (0,204) & (0,233) \end{matrix}$$

$$R^2 = 0,190, BG = 24 \cdot 0,190 = 4,560 > \chi_{\alpha=0,05}^2(1) = 3,84,$$

следовательно, автокорреляция первого порядка присутствует.

Для авторегрессии второго порядка

$$e_t = 0,014 - 0,002 \cdot X_t + 0,289 \cdot e_{t-1} + 0,379 \cdot e_{t-2} + \hat{u}_t, \quad \begin{matrix} (0,167) & (0,102) & (0,217) & (0,217) & (0,223) \end{matrix}$$

$$R^2 = 0,297, BG = 24 \cdot 0,297 = 7,123 > \chi_{\alpha=0,05}^2(2) = 5,99,$$

следовательно, автокорреляция второго порядка присутствует.

По аналогии проверим наличие автокорреляции третьего порядка:

$$e_t = 0,015 - 0,001 \cdot X_t + 0,266 \cdot e_{t-1} + 0,359 \cdot e_{t-2} + 0,089 \cdot e_{t-3} + \hat{u}_t, \quad \begin{matrix} (0,171) & (0,105) & (0,234) & (0,231) & (0,271) & (0,228) \end{matrix}$$

$$R^2 = 0,301, BG = 24 \cdot 0,301 = 7,217 < \chi_{\alpha=0,05}^2(3) = 7,81$$

— автокорреляция третьего порядка отсутствует.

## 12. Автокорреляция: определение, причины, последствия. Н – тест и особенности его применения.

Автокорреляция - это наличие сильной корреляционной зависимости между последовательными наблюдениями одного ряда.

Виды автокорреляции:

- положительная (на графике проявляется чередованием зон положительных и отрицательных остатков).
- отрицательная (остатки "слишком часто" меняются).

Основными причинами автокорреляции являются:

- Спецификации (неправильный выбор формы модели).
- Инерция.
- Эффект "Паутины" (реакция экономических показателей на изменение условий с запазданием).

Последствия автокорреляции:

- Автокорреляция не приводит к смещению оценок регрессии, но оценки перестают быть эффективными.
- Автокорреляция (особенно положительная) часто приводит к уменьшению стандартных ошибок коэффициентов, что влечет за собой увеличение t-статистик
- Оценка дисперсии остатков является смещенной оценкой истинного значения.

В силу вышесказанного выводы по оценке качества коэффициентов и модели в целом, возможно, будут неверными. Это приводит к ухудшению прогнозных качеств модели.

Является модификацией **теста** Дарбина-Уотсона, можно применять для модели с лаговыми переменными. Случайные ошибки должны быть распределены нормально и не должны быть подвержены гетероскедастичности; свободный член не должен быть равен нулю.

Сначала вычисляется статистика Дарбина-Уотсона:

$$DW' = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{ESS} \approx 2(1 - \hat{\rho}), \quad (3.1)$$

### Тест Дарбина (h –тест)

Для обнаружения автокорреляции возмущений в авторегрессионных моделях используется **h**-статистика:

$$h = \tilde{\rho} \sqrt{\frac{n}{1 - n \cdot \tilde{D}(\tilde{\lambda})}},$$

где  $n$  – объём выборки,  $\tilde{D}(\tilde{\lambda})$  – оценка дисперсии коэффициента при лаговой переменной  $Y_{t-1}$ ,

$\tilde{\rho} \approx 1 - \frac{DW}{2}$  – оценка коэффициента автокорреляции первого порядка

### Особенности применения h –теста

1. Вне зависимости от того, сколько лагов переменной  $Y$  включено в модель, значение **h** вычисляется с использованием оценки дисперсии коэффициента при  $Y_{t-1}$
2. **h**-статистика не может быть вычислена при  $n \cdot \tilde{D}(\tilde{\lambda}) > 1$
3. Применение **h**-статистики целесообразно лишь при достаточно большом объёме выборки **n**

При большом объёме **n** и справедливости гипотезы  $H_0: \rho = 0$  **h**-статистика имеет стандартное нормальное распределение,  $h \sim N(0, 1)$ , поэтому если при заданном уровне значимости  $|h| > h_{кр}$ , гипотеза  $H_0$  отклоняется

### 13. Автокорреляция: определение, причины, последствия. Метод рядов Сведа-Эйзенхарта и особенности его применения.

Автокорреляция - это наличие сильной корреляционной зависимости между последовательными наблюдениями одного ряда.

Виды автокорреляции:

- положительная (на графике проявляется чередованием зон положительных и отрицательных остатков).
- отрицательная (остатки "слишком часто" меняются).

Основными причинами автокорреляции являются:

- Спецификации (неправильный выбор формы модели).
- Инерция.
- Эффект "Паутины" (реакция экономических показателей на изменение условий с запазданием).

Последствия автокорреляции:

- Автокорреляция не приводит к смещению оценок регрессии, но оценки перестают быть эффективными.
- Автокорреляция (особенно положительная) часто приводит к уменьшению стандартных ошибок коэффициентов, что влечет за собой увеличение t-статистик
- Оценка дисперсии остатков является смещенной оценкой истинного значения.

В силу вышесказанного выводы по оценке качества коэффициентов и модели в целом, возможно, будут неверными. Это приводит к ухудшению прогнозных качеств модели.

В методе необходимо подсчитать количество положительных и отрицательных отклонений, а также выделить в ряду отклонений подряды последовательных отношений имеющих один знак. Количество таких подрядов обозначить "к".

При использовании метода рядов последовательно определяются знаки отклонений  $e_t$ . Ряд определяется как непрерывная последовательность одинаковых знаков. Количество знаков в ряду называется длиной ряда.

Визуальное распределение знаков свидетельствует о неслучайном характере связей между отклонениями. Если рядов слишком мало по сравнению с количеством наблюдений  $n$ , то вполне вероятно положительная автокорреляция. Если же рядов слишком много, то вероятно отрицательная автокорреляция.

Для более детального анализа предлагается следующая процедура. Пусть:

- $n$  - объем выборки;
- $n_1$  - общее количество знаков "+" при  $n$  наблюдениях (количество положительных отклонений  $e_t$ );
- $n_2$  - общее количество знаков "-" при  $n$  наблюдениях (количество отрицательных отклонений  $e_t$ );
- $k$  - количество рядов.

При достаточно большом количестве наблюдений ( $n_1 > 10$ ,  $n_2 > 10$ ) и отсутствии автокорреляции мы имеем асимптотически нормальное распределение с

$$M(k) = \frac{2n_1n_2}{n_1 + n_2} + 1; \quad D(k) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}.$$

Тогда, если:

$$M(k) - u_{\alpha/2} \cdot D(k) < k < M(k) + u_{\alpha/2} \cdot D(k),$$

то гипотеза об отсутствии автокорреляции не отклоняется.

При небольшом числе наблюдений ( $n_1 > 20$ ,  $n_2 > 20$ ) Свед и Эйзенхарт разработали таблицы критических значений количества рядов при  $n$  наблюдениях.

Суть таблиц заключается в следующем.

На пересечении строки  $n_1$  и столбца  $n_2$  определяются нижнее  $k_1$  и верхнее  $k_2$  значения при уровне значимости 5%.

Если  $k_1 < k < k_2$ , то говорят об отсутствии автокорреляции.

Если  $k \leq k_1$ , то говорят о положительной автокорреляции остатков.

Если  $k \geq k_2$ , то говорят об отрицательной автокорреляции остатков,

14. Модель с автокорреляцией случайного возмущения. Оценка моделей с авторегрессией.

$$y_t = a_0 + a_1x_{1t} + \dots + a_kx_{kt} + \varepsilon_t$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + e_t$$

$\rho$  – коэффициент авторегрессии,  $-1 < \rho < 1$

$\rho > 0$  – положительная автокорреляция

$\rho < 0$  – отрицательная автокорреляция

$\rho = 0$  – автокорреляции нет

## Оценка моделей с авторегрессией

Преобразование переменных:

$$y_t = a_0 + a_1x_{1t} + \dots + a_kx_{kt} + \varepsilon_t$$

$$\rho y_{t-1} = \rho a_0 + a_1\rho x_{1t-1} + \dots + a_k\rho x_{kt-1} + \rho\varepsilon_{t-1}$$

$$y_t - \rho y_{t-1} = a_0(1 - \rho) + a_1(x_{1t} - \rho x_{1t-1}) + \dots$$
$$\dots + a_k(x_{kt} - \rho x_{kt-1}) + \varepsilon_t - \rho\varepsilon_{t-1}$$

$$Y_t = \alpha_0 + a_1X_1 + \dots + a_kX_k + e_t$$

## 15. Процедура Кохрейна-Оркатта.

Используется для исправления автокорреляции. На примере парной регрессионной модели с автокоррелированным возмущением

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

по выборочным данным выполняется настройка модели, и вычисляется вектор остатков регрессии  $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$

по остаткам регрессии оценивается модель авторегрессии первого порядка

$$\epsilon_t = \rho \epsilon_{t-1} + v_t$$

с оценкой  $\rho$  выполняются преобразования переменных:

$$Y_t^* = Y_t - \hat{\rho} Y_{t-1}, X_t^* = X_t - \hat{\rho} X_{t-1}, t = 2, \dots, n$$

по скорректированным выборочным данным определяются МНК оценки параметров  $\alpha^*$   $\beta^*$ . Оценка параметра  $\beta^*$  непосредственно используется в исходной модели,  $\hat{\beta} = \hat{\beta}^*$  оценка параметра вычисляется по формуле

$$\hat{\alpha} = \frac{\hat{\alpha}^*}{1 - \hat{\rho}}$$

По оценкам параметров  $\alpha$  и  $\beta$  вычисляется оценка эндогенной переменной

$$\tilde{Y}_t = \hat{\alpha} + \hat{\beta} X_t$$

5) строится новый вектор остатков, и процедура повторяется с п.2.

Итерационный процесс заканчивается при условии совпадения оценок на последней и предпоследней итерациях с заданной степенью точности.

Оценка (прогноз) эндогенной переменной  $Y$ , в рамках метода Кохрейна-Оркатта, выполняется по формуле

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t + \hat{\rho} \hat{e}_{t-1} = \tilde{Y}_t + \hat{\rho} \hat{e}_{t-1}$$

## 16 Процедура Хилдрета – Лу.

Используется для исправления автокорреляции.

1. Значения  $\rho$  выбираются из интервала  $(-1;1)$  с некоторым шагом.
2. Для каждого значения  $\rho$  выполняются преобразования:

$$\tilde{y}_t = y_t - \rho y_{t-1} \quad \tilde{x}_t = x_t - \rho x_{t-1}$$
$$\tilde{y}_1 = \sqrt{1 - \rho^2} y_1, \quad \tilde{x}_1 = \sqrt{1 - \rho^2} x_1$$

для сохранения первых наблюдений. Определяются МНК-оценки параметров

3. Строится вектор остатков
4. В качестве  $\rho$  выбирается то его значение, для которого сумма квадратов остатков минимальна.
5. В некоторой окрестности полученного значения коэффициента  $\rho$  устраивается более мягкая сетка, и процесс повторяется до тех пор, пока не будет достигнута требуемая точность.

## 17. Оценка влияния факторов, включенных в модель. Коэффициент эластичности, Бета-коэффициент, Дельта – коэффициент.

### Коэффициент эластичности.

Показывает, на сколько процентов изменится зависимая переменная при изменении j-го фактора на 1%.

$$\mathcal{E}_j = \hat{a}_j \cdot \frac{\bar{x}_j}{\bar{y}} \approx \frac{\delta y}{\bar{y}} / \frac{\delta x}{\bar{x}}$$

Эластичность не нормирована и может изменяться от  $-\infty$  до  $+\infty$ . Высокий уровень эластичности означает сильное влияние независимой переменной на объясняемую переменную.

Однако средний частный коэффициент эластичности *не учитывает степени колеблемости факторов*, которая может значительно различаться у отдельных факторов. Поэтому для устранения различий в измерении и степени колеблемости факторов используется другой показатель.

Бета-коэффициент - коэффициент регрессии в стандартизированном масштабе:

$$\beta_j = \hat{\alpha}_j \cdot \frac{S_{x_j}}{S_y}$$

где  $S_{x_j}$  — среднее квадратическое отклонение фактора  $j$ .

$$\text{где } S_{x_j}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \quad S_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2.$$

Бета-коэффициент показывает, на какую часть величины среднего квадратического отклонения изменяется среднее значение зависимой переменной с изменением соответствующей независимой переменной на одно среднее квадратическое отклонение при фиксированном на постоянном уровне значения остальных независимых переменных.

### Дельта-коэффициент

Дельта - коэффициент показывает долю влияния фактора в суммарном влиянии всех факторов:

$$\Delta_j = r_{y,x_j} \cdot \beta_j / R^2$$

где  $r_{y,x_j}$  — коэффициент парной корреляции между фактором  $j$  и зависимой переменной.

В практических задачах при корректно проведенном анализе величины дельта - коэффициенты положительны, то есть все коэффициенты регрессии имеют тот же знак, что и соответствующие парные коэффициенты корреляции. Указанные характеристики позволяют упорядочить факторы по степени влияния факторов на зависимую переменную.

Коэффициент эластичности	$\mathcal{E}_j = \hat{a}_j \cdot \frac{\bar{x}_j}{\bar{y}}$	Показывает, на сколько процентов изменится зависимая переменная при изменении j-го фактора на 1%. Высокий уровень эластичности означает сильное влияние независимой переменной на объясняемую переменную.
Бета-коэффициент	$\beta_j = \hat{\alpha}_j \cdot \frac{S_{x_j}}{S_y}$	На какую часть своего СКО изменится значение исследуемой переменной при изменении соответствующего фактора на 1 СКО. Используется для устранения различий в измерении и степени колеблемости факторов. Вспомогательный, используется редко
Дельта-коэффициент	$\Delta_j = r_{y,x_j} \cdot \beta_j / R^2$	Показывает среднюю долю влияния соответствующего фактора в совокупном влиянии всех факторов, включенных в модель. В сумме дает 1.

## 18. Мультиколлинеарность: понятие, причины и последствия

Под мультиколлинеарностью понимается тесная взаимная коррелированность объясняющих переменных  $X_j$ . Мультиколлинеарность может проявляться в функциональной (явной) и стохастической (скрытой) формах. При функциональной форме мультиколлинеарности по крайней мере одна из парных связей между объясняющими переменными  $X_j$  и  $X_k$  является линейной функциональной зависимостью  $r_{X_j, X_k}$ . В этом случае матрица  $(X^T X)$  является особенной, так как содержит линейно зависимые векторы-столбцы и ее определитель равен нулю, что приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели.

В экономических исследованиях мультиколлинеарность чаще проявляется в стохастической форме, когда между хотя бы двумя объясняющими переменными  $X_j$  и  $X_k$  существует тесная корреляционная связь. Матрица  $(X^T X)$  в этом случае не является особенной, но ее определитель очень мал.

### Причины мультиколлинеарности:

1. наличие причинно-следственных связей между социально-экономическими явлениями.
2. транзитивность

### Последствия:

Оценки полученные по МНК остаются несмещенными и эффективными, но перестают быть состоятельными

1. Большая дисперсия оценок, приводят к расширению интервала и снижению точности статистики полученных на основе их расчета
2. Затрудняется определение вклада каждой из объясняющих переменных в объясняемую переменную уравнение регрессии дисперсию зависимой переменной
3. Неверные знаки у коэффициентов регрессии
4. Незначимость коэффициентов регрессии в большем количестве при значимости уравнения в целом
5. Значительные изменения коэффициентов регрессии при изменении состава экзогенных переменных и объектов входящих в выборку

## 19. Выявление мультиколлинеарности: коэффициент увеличения дисперсии (VIF –тест)

Еще один метод измерения мультиколлинеарности является следствием анализа формулы стандартной ошибки коэффициента регрессии:

$$s_{a_j} = \frac{\sigma_y}{\sigma_{x_j}} \sqrt{\frac{1 - R_{yx_1 \dots x_p}^2}{(1 - R_{x_j x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2)(n - m - 1)}}$$

Как следует из данной формулы, стандартная ошибка будет тем больше, чем меньше будет величина, которую называют фактор инфляции дисперсии (или фактор вздутия дисперсии) VIF:

$$VIF_{x_j} = \frac{1}{(1 - R_{x_j x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2)}$$

где  $R_{x_j x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2$  — коэффициент детерминации, найденный для уравнения зависимости переменной  $x_j$  от других переменных  $x_1 \dots x_p$ , входящих в рассматриваемую модель множественной регрессии. Так как величина  $R_{x_j x_1 \dots x_{j-1} x_{j+1} \dots x_p}^2$  отражает тесноту связи между переменной  $x_j$  и прочими объясняющими переменными, то она, по сути, характеризует мультиколлинеарность применительно к данной переменной  $x_j$ . При отсутствии связи показатель  $VIF_{x_j}$  будет равен (или близок) единице, усиление связи ведет к стремлению этого показателя к бесконечности. Считают, что если  $VIF_{x_j} > 3$  для каждой переменной  $X_j$ , то имеет место мультиколлинеарность.

## 20. Выявление мультиколлинеарности: Алгоритм Фаррара-Глобера

Можно проводить проверку наличия мультиколлинеарности всего массива данных, используя статистику Фаррара-Глоубера, по формуле:

$$FG_{\text{набл}} = -\left(n - 1 - \frac{1}{6}(2m + 5)\right) \ln(\det R),$$

где  $n$ -число наблюдений; $m$ -число объясняющих факторов;  $R$ -матрицы парных коэффициентов корреляции

Фактическое значение этого критерия сравнивается с табличным(критическим) значением с использованием статистики  $\chi^2$ . Значение  $\chi^2(\alpha, v)$  зависит от уровня значимости  $\alpha$  и числа степеней свободы  $v = \frac{1}{2}m(m - 1)$ , здесь  $m$ -число объясняющих переменных.Критическое значение можно найти при помощи функции ХИ2ОБР MS EXcel

## 21. Построение гребневой регрессии. Суть регуляризации.

Применение ридж-регрессии (гребневой регрессии) предполагает корректировку элементов главной диагонали матрицы  $(X^T X)$  на некую произвольно задаваемую положительную величину  $\tau$ . Значение рекомендуется брать от 0,1 до 0,4. Дрейпер, Смит в своей работе приводят один из способов "автоматического" выбора величины  $\tau$ , предложенный Хоэрлом, Кеннардом и Белдвинном:

$$\tau = \frac{mSS_e}{n-m-1a^{*T}a^*} \quad (1)$$

где  $m$  — количество параметров (без учета свободного члена) в исходной модели регрессии;  $SS_e$  — остаточная сумма квадратов, полученная по исходной модели регрессии без корректировки на мультиколлинеарность;  $a^*$  — вектор-столбец коэффициентов регрессии, преобразованных по формуле:

$$a_j^* = a_j * \sqrt{\sum (x_j - \bar{x}_j)^2} \quad (2)$$

где  $a_j$  — параметр при переменной  $x_j$  в исходной модели регрессии.

После выбора величины  $\tau$  формула для оценки параметров регрессии будет иметь вид:

$$a_\tau = (X_\tau^T X_\tau + \tau I)^{-1} X_\tau^T Y_\tau \quad (3)$$

где  $I$  — единичная матрица;  $X_\tau$  — матрица значений независимых переменных: исходных или преобразованных по формуле 4;  $Y_\tau$  — вектор значений зависимой переменной: исходных или преобразованных по формуле (5).

При построении ридж-регрессии рекомендуется преобразовывать независимые переменные:

$$x_{\tau j} = \frac{x_j - \bar{x}_j}{\sqrt{\sum (x_j - \bar{x}_j)^2}} \quad (4)$$

и результирующую переменную

$$y_\tau = y - \bar{y} \quad (5)$$

В этом случае после оценки параметров по формуле (3) необходимо перейти к регрессии по исходным переменным, используя соотношения

$$a_j = \frac{a_{\tau j}}{\sqrt{\sum (x_j - \bar{x}_j)^2}}, \quad j = 1, 2, \dots, p; \quad a_0 = \bar{y} - \sum_j a_j \bar{x}_j \quad (6)$$

Оценки параметров регрессии, полученные с помощью формулы (3) будут смещенными. Однако так как определитель матрицы  $(X^T X + \tau I)$  больше определителя матрицы  $(X^T X)$ , дисперсия оценок параметров регрессии уменьшится, что положительно повлияет на прогнозные свойства модели.

## 22. Алгоритм пошаговой регрессии

Абсолютно надёжным методом поиска наилучшего состава регрессоров из списка является перебор всех возможных комбинаций. Существует несколько вариантов направленности алгоритма

1. Включения (каждую итерацию включается по одному фактору)
2. Исключения (каждую итерацию исключается факторы)
3. Двухнаправленный (сочетание 1 и 2 методов)

Целевой функцией для оптимизации является  $R^2$  всей модели и  $t$ -статистика для каждого из факторов

## 23. Метод главных компонент (РСА) как радикальный метод борьбы с мультиколлинеарностью

Применение метода главных компонент предполагает переход от взаимозависимых переменных  $\mathbf{x}$  к независимым друг от друга переменным  $\mathbf{z}$ , которые называют **главными компонентами**. Каждая главная компонента  $z_j$  может быть представлена как линейная комбинация центрированных объясняющих переменных  $t_j$ . Центрирование переменной:

$$t_{ji}^* = x_{ji} - \bar{x}_j \quad (1)$$

Стандартизация(масштабирование):

$$t_{ji} = \frac{x_{ji} - \bar{x}_j}{\sigma_{x_j}} \quad (2)$$

Количество компонент может быть меньше или равно количеству исходных независимых переменных  $\mathbf{p}$ . Компоненту с номером  $\mathbf{k}$  можно записать следующим образом:

$$z_k = f_{k1}t_1 + f_{k2}t_2 + \dots + f_{kp}t_p \quad (3)$$

Доля дисперсии  $k$ -й компоненты в общей дисперсии независимых переменных рассчитывается по формуле:

$$d_k = \frac{\lambda_k}{\sum_j \lambda_j} \quad (4)$$

где  $\lambda_k$  — собственное число, соответствующее данной компоненте; в знаменателе формулы (4) приведена сумма всех собственных чисел матрицы  $\frac{1}{n}(T^T T)$ , где  $T$ -матрица размером  $(n * p)$ , содержащая стандартизированные переменные.

После расчета значений компонент  $z_j$  строят регрессию, используя МНК. Зависимую переменную в регрессии по главным компонентам (5) целесообразно центрировать (стандартизовать) по формулам (1) или (2).

$$t_y = b_1 z_1 + b_2 z_2 + \dots + b_k z_k + \delta \quad (5)$$

где  $t_y$  — стандартизованная (центрированная) зависимая переменная;  $b_1, b_2, \dots, b_k$  — коэффициенты регрессии по главным компонентам;  $z_1, z_2, \dots, z_k$  — главные компоненты, упорядоченные по убыванию собственных чисел  $\lambda_k$ ;  $\delta$  -случайный остаток.

После оценки параметров регрессии (5) можно перейти к уравнению регрессии в исходных переменных, используя выражения (1)—(3).

## 24. Фиктивная переменная и правило её использования

Фиктивная переменная — качественная переменная, принимающая значения 0 и 1, включаемая в эконометрическую модель для учёта влияния качественных признаков и событий на объясняемую переменную.

Пусть имеется признак, который принимает несколько возможных значений. Общее правило введения фиктивных переменных следующее: общее количество фиктивных переменных должно быть на единицу меньше количества возможных значений качественного признака, если в модели имеется константа. Это необходимо, чтобы не возникла проблема полной коллинеарности переменных.

Например, уровень образования: нет образования, среднее образование, высшее образование, ученая степень и т. д. В этом случае каждому уровню образования, кроме уровня «нет образования» можно поставить в соответствие некоторую фиктивную переменную.

## 25. Модель дисперсионного анализа

Дисперсионный анализ — метод, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. В отличие от t-критерия, позволяет сравнивать средние значения трёх и более групп.

В зависимости от типа и количества переменных различают:

- однофакторный и многофакторный дисперсионный анализ (одна или несколько независимых переменных);
- одномерный и многомерный дисперсионный анализ (одна или несколько зависимых переменных);
- дисперсионный анализ с повторными измерениями (для зависимых выборок);
- дисперсионный анализ с постоянными факторами, случайными факторами, и смешанные модели с факторами обоих типов;

Ход дисперсионного анализа:

- 1) Подсчитать значение  $SS_{\text{факт}}$  - означает вариативность признака, обусловленную действием исследуемого фактора

$$SS_{\text{факт}} = \frac{\sum \sum x_{ij}^2}{n} - \frac{(\sum x_i)^2}{N}$$

N - общее количество наблюдений

- 2) Подсчитать  $SS_{\text{общ}}$  - общую вариативность признака.

$$SS_{\text{общ}} = \sum x_i^2 - \frac{(\sum x_i)^2}{N}$$

- 3) Подсчитать случайную (остаточную) величину  $SS_{\text{сл}} = SS_{\text{общ}} - SS_{\text{факт}}$

- 4) Определить число степеней свободы:

$$df_{\text{факт}} = \text{кол-во объясняющих переменных} - 1$$

$$df_{\text{общ}} = \text{кол-во всех наблюдений} - 1$$

$$df_{\text{сл}} = df_{\text{общ}} - df_{\text{факт}}$$

- 5) Разделить каждую SS на соответствующую степень свободы

- 6) Посчитать значение  $F_{\text{расч}} = \frac{MS_{\text{факт}}}{MS_{\text{сл}}}$  и сравнить его с табличным

## 26. Модель ковариационного анализа

Ковариационный анализ – методы изучения взаимосвязи между количественной зависимой переменной и набором категориальных и одновременно набором количественных предикторов. Независимые количественные предикторы в модели ковариационного анализа называют ковариатами, а категориальные независимые переменные – факторами.

Ковариационный анализ является как бы синтезом регрессионного и дисперсионного анализа. Основные теоретические и прикладные проблемы ковариационного анализа относятся к линейным моделям. Если в линейной модели взаимосвязи присутствуют только категориальные предикторы с помощью введения фиктивных переменных, то получается модель дисперсионного анализа. Если в линейной модели присутствуют только количественные предикторы – получается модель регрессионного анализа. А при совместном введении факторов и ковариат проводится ковариационный анализ.

По отношению к зависимой переменной ковариаты являются сопутствующими переменными. Ковариационный анализ часто используется при «управлении» эффектами внешних переменных. Другими словами, введение в модель ковариат позволяет оценить их влияние на взаимодействие зависимой переменной и факторов. Например, аналитик может использовать коэффициент IQ студентов в качестве ковариаты (количественный предиктор) при исследовании эффективности различных методов обучения (качественный предиктор).

---

## 27. Фиктивные переменные в сезонном анализе

Было бы здорово в процессе экономического моделирования учитывать влияние сезонной составляющей => можно использовать фиктивные переменные

Фиктивные переменные - которые принимают значение 0 или 1

0 - отсутствие признака в наблюдении

1 - наличие признака в наблюдении

Моделирование сезонных переменных

Для этого воспользуемся следующими фиктивными переменными:

•  $d_1$  - 1 для первого квартала, 0 для других

•  $d_2$  - 1 для второго квартала, 0 для других

•  $d_3$  - 1 для третьего квартала, 0 для других

Состояние где фиктивные переменные нулевые - базовое

спецификация модели сдвига:

$$y_i = \beta_1 + \beta_2 x_i + \beta_{d_1} d_1 i + \beta_{d_2} d_2 i + \beta_{d_3} d_3 i + E_i$$

$$E(y_i | d_1 = 1) = (\beta_1 + \beta_{d_1}) + \beta_2 x_i$$

$$E(y_i | d_2 = 1) = (\beta_1 + \beta_{d_2}) + \beta_2 x_i$$

$$E(y_i | d_3 = 1) = (\beta_1 + \beta_{d_3}) + \beta_2 x_i$$

$$E(y_i | d_1, d_2, d_3 = 0) = \beta_1 + \beta_2 x_i$$

Экономический смысл коэффициентов заключается в том, насколько в среднем изменится  $y_i$  при фиксированных значениях регрессора при переходе из сезона в сезон

## 28. Фиктивная переменная сдвига: спецификация регрессионной модели с фиктивной переменной сдвига; экономический смысл параметра при фиктивной переменной; смысл названия

Было бы здорово в процессе эконометрического моделирования изучать и учитывать не только количественные но и качественные переменные => можно использовать фиктивные переменные.

Фиктивные переменные - которые принимают значение 0 или 1

0 - отсутствие признака в наблюдении

1 - наличие признака в наблюдении

Спецификация модели с фиктивной переменной сдвига:

$$y_t = \beta_1 + \beta_2 x_t + \beta_d d_t + E_t$$

$x_t$  - количественная переменная;  $d_t$  - качественная характеристика

Экономическая характеристика:

$$E(y_t | d_t = 0) = \beta_1 + \beta_2 \cdot x_t$$

$$E(y_t | d_t = 1) = (\beta_1 + \beta_d) + \beta_2 x_t$$

$\beta_d$  (параметр при фиктивной переменной сдвига) - среднее изменение изучаемого признака  $y$  при переходе из одной категории в другую при неизменяемых значениях остальных параметров. Влияет только на изменение свободного члена в уравнении регрессии, поэтому и называется сдвигом.

## 29. Фиктивная переменная наклона: спецификация регрессионной модели с фиктивной переменной наклона; экономический смысл параметра при фиктивной переменной; смысл названия

Фиктивная переменная наклона изменяет наклон линии регрессии.

Фиктивные переменные - которые принимают значение 0 или 1

0 - отсутствие признака в наблюдении

1 - наличие признака в наблюдении

Фиктивные переменные наклона дают возможность построить линейные модели переменного наклона (кусочно-линейные), позволяющие учесть структурные изменения в экономических процессах.

Спецификация модели с фиктивной переменной наклона:

$$y_i = \beta_0 + \beta_1 x + \beta_2 \gamma x + E$$

### 30. Определение структурных изменений в экономике: использование фиктивных переменных, тест Чоу

В экономике структурные изменения — это сдвиг или изменение основных способов функционирования рынка или экономики.

Структурные изменения в экономике могут быть вызваны следующими факторами:

- экономическое развитие страны
- глобальные сдвиги в капитале и рабочей силе
- изменения в доступности ресурсов вследствие войны или стихийных бедствий, открытия или истощения природных ресурсов
- изменений в политической системе

Фиктивная переменная — качественная переменная, принимающая значения 0 и 1, включаемая в эконометрическую модель для учёта влияния качественных признаков и событий на объясняемую переменную.

Тест Чоу - это процедура проверки стабильности параметров регрессионной модели, наличия структурных сдвигов в выборке. Фактически тест проверяет неоднородность выборки в контексте регрессионной модели.

Модель для первого набора наблюдений:

$$Y = \beta'_0 + \beta'_1 X_1 + \dots + \beta'_k X_k + \epsilon'$$

Модель для второго набора наблюдений:

$$Y = \beta''_0 + \beta''_1 X_1 + \dots + \beta''_k X_k + \epsilon''$$
$$H_0 : \beta'_0 = \beta''_0, \dots, \beta'_k = \beta''_k, \sigma_{\epsilon'}^2 = \sigma_{\epsilon''}^2$$
$$H_1 : \exists i : \beta'_i \neq \beta''_i$$

Тестовая статистика в тесте Чоу:

$$F = \frac{(RSS_R - RSS_{UR}) / (k+1)}{(RSS_1 + RSS_2) / (n-2(k+1))} = \frac{(RSS_p - (RSS_1 + RSS_2)) / (k+1)}{(RSS_1 + RSS_2) / (n-2(k+1))}$$

k - кол-во всех регрессоров

$RSS_p$  - это сумма квадратов остатков для всей выборки

$RSS_1$  - это сумма квадратов остатков для выборки 1

$RSS_2$  - это сумма квадратов остатков для выборки 2

Если  $F > F_{\text{критическое}}$  (при выбранном уровне значимости), то основная гипотеза отвергается и нужно оценивать две отдельные регрессии.

### 31. Модели бинарного выбора. Недостатки линейной модели.

Зависимая переменная принимает только два значения, обычно 0 и 1.

Такие переменные называются *бинарными*.

Приведем некоторые примеры.

1. После окончания школы выпускник решает: пойти учиться в вуз ( $Y = 1$ ) или нет ( $Y = 0$ ). 2. Владелец фирмы решает: инвестировать средства в определенный проект ( $Y = 1$ ) или не инвестировать ( $Y = 0$ ).

В этом случае, как и в предыдущих, можно использовать линейную модель

$$y = \beta_0 + \beta_1 * x_1 + \dots + \beta_i * x_i + \epsilon_i$$

Тогда  $p = p(Y = 1), i = 1, \dots, n$

С другой стороны,  $p(Y = 0) = 1 - p$

$$E(Y) = p(Y = 1) * 1 + P(Y = 0) * 0$$

Таким образом, оценивая модель с зависимой переменной, принимающей два значения, мы оцениваем вероятность того, что эта переменная принимает значение 1.

Если мы предполагаем, что зависимость от выбранных факторов  $X_1, \dots, X_k$  является линейной, то получаем модель линейной вероятности:

$$P(Y = 1) = \beta_0 + \beta_1 * x_1 + \dots + \beta_k * x_k$$

Главный недостаток этой модели очевиден: будучи линейной функцией от объясняющих факторов, оцененное значение вероятности запросто может не принадлежать отрезку  $[0; 1]$ .

В классической линейной регрессионной модели обычно вводится предположение о нормальности распределения ошибок. Это дает возможность, в частности, использовать t-статистики для проверки гипотез о значимости коэффициентов. Однако если  $Y$ , принимает всего два значения, то и ошибки принимают только два значения:  $1 - (p_0 + p_1 x_1 + \dots + p_k x_k)$  с вероятностью  $p_x$  и  $(p_0 + p_1 x_1 + \dots + p_k x_k)$  с вероятностью  $1 - p_x$ . Это распределение никак нельзя считать нормальным, а следовательно, нельзя проводить привычные тесты.

Так имеет место гетероскедастичность, в этом случае МНК-оценки не являются эффективными.

### 32. Модели множественного выбора: модели с неупорядоченными альтернативными вариантами

В данных моделях альтернативы не упорядочены, и выбор осуществляется на основе функции полезности. Варианты множественного выбора нумеруются в произвольном порядке 1, 2, 3, ..., J. Пусть  $i$ -ый субъект исследования приписывает  $j$ -ой альтернативе некоторую случайную полезность

$$u_{ij} = x_{ij} * \beta + \epsilon_{ij}, i = 1, \dots, n, j = 1, \dots, J$$

Поскольку *критерием отбора* альтернативы является *полезность*, то выбирается такая альтернатива, которая её максимизирует:

В моделях множественного выбора с неупорядоченными альтернативами, для упрощения, принимаются следующие предпосылки: либо постоянными являются параметры, а специфичными по отношению к альтернативам являются регрессоры; либо регрессоры для всех альтернатив одни и те же, а специфичными являются параметры модели.

$$P(y_i = j) = \frac{\{\exp(x_{ij} * \beta - x_{i1} * \beta)\}}{1 + \exp(x_{i2} * \beta - x_{i1} * \beta) + \dots + \exp(x_{iJ} * \beta - x_{i1} * \beta)}$$

формула для определения вероятности альтернативы, известная как мультиномиальная логит-модель (multinomial logit model) (множественная логит-модель)

$$P(y_i = j) = \frac{(\exp(x_{ij} * \beta))}{(1 + \exp(x_{i2} * \beta) + \dots + \exp(x_{iJ} * \beta))}$$

Во втором случае, когда специфическими по отношению к альтернативам являются параметры, мультиномиальная модель принимает вид:

$$P(y_i = j) = \frac{\exp(x_i * \beta^j)}{1 + \exp(x_i * \beta^2) + \dots + \exp(x_i * \beta^J)}$$

Логит-модель бинарного выбора является частным случаем модели множественного выбора при  $J = 2$ . Для оценки параметров модели применяется метод максимального правдоподобия.

### 33. Модели множественного выбора: модели с упорядоченными альтернативными вариантами

Обоснование модели множественного выбора с упорядоченными альтернативами выполняется в рамках порогового подхода, в котором используется несколько пороговых значений ненаблюдаемой переменной. Модель строится на основе латентной переменной, линейно зависящей от вектора объясняющих переменных

$$y = \begin{cases} y^* \leq u_1 \\ u_1 < y^* \leq u_2 \dots \\ y^* > u_{J-1} \end{cases}$$

где уровни цензурирования (пороговые значения для эндогенной переменной) либо известны, либо неизвестны и оцениваются вместе с параметрами

Вероятности альтернатив определяются по формуле

$$p(y_i = j) = F(\mu_j - x\beta) - F(\mu_{j-1} - x\beta), j = 1, \dots, J$$

где  $F(-)$  — функция распределения случайного возмущения модели.

В зависимости от выбора модели случайные возмущения имеют логистическое или нормальное распределение.

### 34. Модели множественного выбора: гнездовые logit-модели

Гнездовые logit-модели, также известные как мультиномиальные logit-модели, являются статистическими моделями, которые используются для анализа множественного выбора с упорядоченными альтернативами. Они представляют собой расширение бинарной logit-модели на случай множественного выбора.

Гнездовая logit-модель предполагает, что вероятность выбора каждой альтернативы зависит от линейной комбинации регрессоров и параметров модели.

В гнездовой logit-модели, помимо собственных характеристик группы, используется их "ценность" для определения вероятности выбора.

"Ценность" группы может быть определена как линейная комбинация ее характеристик с соответствующими параметрами модели.

Предполагая, что у нас есть  $L$  групп (альтернатив) и каждая группа имеет свои характеристики, обозначенные как  $X = (X_1, X_2, \dots, X_k)$ , где  $k$  - общее количество характеристик для каждой группы, "ценность" группы  $l$  может быть определена следующим образом:

$$L = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

где  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  - параметры модели, которые определяют вклад каждой характеристики в "ценность" группы  $l$ .

Затем, вероятность выбора  $l$ -й группы (альтернативы) для данного наблюдения  $i$  может быть определена с использованием

$$p(y_i = l | X_i) = \frac{\exp(L)}{\sum_{j=1}^L (\exp(j))}$$

где  $p(y_i = l | X_i)$  - вероятность выбора  $l$ -й группы (альтернативы) для наблюдения  $i$ ,  $X_i$  - характеристики наблюдения  $i$ ,  $L$  - общее количество групп (альтернатив).

Таким образом, гнездовая logit-модель учитывает не только характеристики каждой группы (альтернативы), но и их "ценность", определенную через линейную комбинацию параметров модели и характеристик. Это позволяет учесть дополнительные факторы или предпочтения, влияющие на выбор конкретной группы (альтернативы) в рамках множественного выбора.

Качество оценок, получаемых на основе гнездовой logit-модели, во многом определяется правильностью построения дерева альтернативных вариантов. Отметим, что на практике достаточно трудно оценить, соответствует ли выбранная структура такого дерева исходным условиям модели, состоящих в постулировании определенных допущений относительно дисперсий ошибок (постоянство дисперсий ошибок внутри группы и различие дисперсий в разных группах).

### 35. Модели счетных данных (отрицательная биномиальная модель, hurdle-model)

Модели счетных данных используются для анализа переменных, которые представляют собой дискретные счетные значения. Отрицательная биномиальная модель является статистической моделью, применяемой для анализа счетных данных. Она является обобщением распределения Пуассона и предполагает, что дисперсия счетных данных превышает их среднее значение, что называется избыточностью дисперсии (overdispersion). Функция вероятности отрицательного биномиального распределения для счетной переменной  $Y$  задается следующим образом:

$$p(Y = y) = \frac{\Gamma(y + \theta)}{\Gamma(\theta) * y!} * \frac{\mu}{(\mu + \theta)^y} * \left(\frac{\theta}{\mu + \theta}\right)^\theta$$

где  $y$  - конкретное значение счетной переменной,  $\Gamma()$  - функция гамма,  $y!$  - факториал  $y$ ,  $\mu$  - среднее значение счетной переменной.  $\theta$  - параметр избыточности дисперсии (обратно пропорционален степени избыточности дисперсии). В моделировании отрицательной биномиальной модели применяется метод максимального правдоподобия для оценки параметров  $\mu$  и  $\theta$  на основе имеющихся данных. Это позволяет определить, какие факторы или переменные оказывают статистически значимое влияние на счетную переменную. Отрицательная биномиальная модель позволяет более точно моделировать и интерпретировать данные счетных переменных, учитывая их специфические особенности, и обеспечивает более надежные статистические выводы. Hurdle-модель состоит из двух компонент: компонента вероятности наличия нуля (zero-inflation component) и компонента счетного значения при отличном от нуля (count component). Компонента вероятности наличия нуля: В этой части модели используется бинарная логистическая регрессия или другая модель для моделирования вероятности наличия нуля. Пусть  $p_i$  будет вероятностью наличия нуля для  $i$ -го наблюдения. Тогда формула для моделирования вероятности наличия нуля выглядит следующим образом:

$$p_i = p(Y_i = 0) = \frac{1}{1 + \exp(-z_i)}$$

где  $z_i$  - линейная комбинация регрессоров и соответствующих коэффициентов, связанных с вероятностью наличия нуля. Компонента счетного значения при отличном от нуля: Эта часть модели используется для моделирования счетного значения только для наблюдений, которые не равны нулю. Пусть  $y_i$  будет счетным значением для  $i$ -го наблюдения, где  $y_i > 0$ . В этой части модели используется распределение Пуассона или отрицательное биномиальное распределение. Формула для моделирования счетного значения отличного от нуля выглядит следующим образом:

$$p(Y_i = y | Y_i > 0) = \frac{\exp(-\mu_i) * \mu_i^y}{y!}$$

где  $\mu_i$  - среднее значение счетного значения для  $i$ -го наблюдения, которое зависит от регрессоров и соответствующих коэффициентов. В hurdle-модели общая вероятность  $P(Y_i = y)$  для каждого наблюдения  $i$  считается как произведение вероятности наличия нуля и вероятности счетного значения отличного от нуля:

$$p(Y_i = y) = p_i * p(Y_i = y | Y_i > 0)$$

где  $p_i$  - вероятность наличия нуля для  $i$ -го наблюдения,  $P(Y_i = y | Y_i > 0)$  - вероятность счетного значения отличного от нуля для  $i$ -го наблюдения. Hurdle-модель широко используется для анализа данных с избыточным количеством нулей в различных областях, таких как медицинская статистика, экономика, социология и экология. Она позволяет более точно моделировать и интерпретировать данные счетных переменных, учитывая их особенности и обеспечивает более надежные статистические выводы.

### 36. Модели усеченных выборок

Выборку называют усеченной, если она производится из части интересующей исследователя генеральной совокупности. Пусть переменная  $y$  принимает значения из множества. Если мы можем наблюдать  $y$  только когда она принадлежит подмножеству тогда  $y$  называется усеченной. Пример правила отбора: только индивиды с  $y > c$  попадают в выборку, например для анализа продаж люкс автомобилей берут только класс богатых людей

Модели усеченных выборок имеют две основные формулировки:

Усечение снизу (left-truncation): В этом случае, зависимая переменная имеет нижнюю границу и значения ниже этой границы не наблюдаемы. Например, в экономическом анализе, если изучаются доходы работников, но данные доступны только для тех, кто имеет доход выше определенного порога, то это будет случай усечения снизу.

$$Y^* = \max(Y, L)$$

$$E(Y^*|X) = x\beta$$

$$\text{Var}(Y^*|X) = \sigma^2$$

Для дискретных данных:

$$P(Y^* = y|X) = \frac{P(Y = y)}{1 - P(Y \leq L)} E(Y^* | X) = \sum_{\{y=L+1\}}^{\infty} y * \frac{(P(Y = y))}{1 - P(Y \leq L)}$$

Усечение сверху (right-truncation): В этом случае, зависимая переменная имеет верхнюю границу и значения выше этой границы не наблюдаемы. Например, при анализе продаж недвижимости, если данные доступны только для продаж с ценой ниже определенного предела, то это будет случай усечения сверху.

Для непрерывных данных:

$$Y^* = \min(Y, L)$$

$$E(Y^*|X) = x\beta$$

$$\text{Var}(Y^*|X) = \sigma^2$$

Для дискретных данных:

$$P(Y^* = y|X) = \frac{P(Y = y)}{1 - P(Y \leq U)}$$
$$E(Y^* | X) = \sum_{\{y=L+1\}}^{\infty} y * \frac{(P(Y = y))}{1 - P(Y \leq U)}$$

### 37. Модели цензурированных выборок (tobit-модель)

Модели цензурированных данных возникают, когда точное значение переменной неизвестно, но известно множество, к которому это значение принадлежит. Оценивание моделей с цензурированными данными с помощью метода наименьших квадратов не даст нам состоятельные оценки параметров, поскольку цензурированная выборка нерепрезентативна для генеральной совокупности. Из моделей цензурированных данных мы рассмотрим модель Тобина, или тобит-модель.

Тобит-модель. Пусть ненаблюдаемая зависимая переменная  $y^*$  удовлетворяет регрессионному уравнению

$$y_i^* = x_i\beta + \epsilon_i$$
$$y_i^* = \begin{cases} y_i^*, & y_i^* \geq L \\ L, & y_i^* < L \end{cases}$$

$L$  - порог цензурирования

Данная спецификация соответствует тобит-модели с цензурированием снизу (или слева).

Вероятность того, что наблюдение будет цензурирована, равна

$$p(y_i^* \leq L) = p(x_i\beta + \epsilon_i \leq L) = \Phi\left(\frac{L - x_i\beta}{\sigma}\right)$$

Выборка также может иметь цензурирование сверху (или справа), тогда наблюдаемый  $y$  будет выглядеть так:

(Поменять знаки в системе

(1) на  $\leq$

(2) на  $>$  и  $L$  на  $U$ )

где  $U$  — порог цензурирования сверху.

Возможна ситуация, когда цензурирование есть и сверху, и снизу одновременно. Более сложной спецификацией является случай, когда неизвестна граница цензурирования и ее тоже требуется оценить.

### 38. Модели случайно усеченных выборок (selection model)

Усеченная выборка возникает, когда значения некоторой переменной наблюдаются только в определенных пределах или в зависимости от некоторого условия.

Примером усеченной выборки может служить ситуация, когда исследователи регистрируют только тех людей, которые имеют определенное свойство, и не регистрируют остальных. Например, при изучении выживаемости пациентов с определенным заболеванием, выборка может быть усечена до только тех пациентов, которые пережили определенное количество времени.

Модели усеченных выборок имеют две основные формулировки:

Усечение снизу (left-truncation): В этом случае, зависимая переменная имеет нижнюю границу и значения ниже этой границы не наблюдаемы. Например, в экономическом анализе, если изучаются доходы работников, но данные доступны только для тех, кто имеет доход выше определенного порога, то это будет случай усечения снизу.

$$Y^* = \max(Y, L)$$

$$E(Y^*|X) = x\beta$$

$$\text{Var}(Y^*|X) = \sigma^2$$

Для дискретных данных:

$$P(Y^* = y|X) = \frac{P(Y = y)}{1 - P(Y \geq L)}$$

$$E(Y^*|X) = \sum_{\{y=L+1\}}^{\infty} y * \frac{(P(Y = y))}{1 - P(Y \geq L)}$$

Усечение сверху (right-truncation): В этом случае, зависимая переменная имеет верхнюю границу и значения выше этой границы не наблюдаемы. Например, при анализе продаж недвижимости, если данные доступны только для продаж с ценой ниже определенного предела, то это будет случай усечения сверху.

Для непрерывных данных:

$$Y^* = \min(Y, L) \quad E(Y^*|X) = x\beta \quad \text{Var}(Y^*|X) = \sigma^2$$

Для дискретных данных:

$$P(Y^* = y|X) = \frac{P(Y = y)}{1 - P(Y \leq U)}$$

$$E(Y^*|X) = \sum_{\{y=L+1\}}^{\infty} y * \frac{(P(Y = y))}{1 - P(Y \leq U)}$$

### 39. Логит-модель. Этапы оценки. Области применения.

Логит модель-модель бинарного выбора, в которой случайные остатки подчиняются логистическому закону распределения.

$$Y_i = \frac{(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}{(1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}))}$$

При любых значениях факторных переменных и коэффициентов регрессии, значения результирующей переменной  $Y_i$  лежат в интервале  $[0; +1]$ .

Обобщенный вид модели, в которой результирующая переменная может произвольно меняться внутри заданного числового интервала

$$Y_i = \frac{\beta_0}{(1 + \beta_1 \exp(\beta_2 X_i))}$$

Этапы

Определение зависимой переменной и факторов

Построение переменной  $Z$ , как линейной комбинации независимых переменных

Построение уравнения для искомой вероятности события и нахождение производных (для оценки кумулятивного и предельного воздействия факторов)

Проведение вычислений с помощью программы (используется метод максимального правдоподобия)

Интерпретация результатов

Качество оценивания

Применяется в медицине (вероятность успешного лечения), социологии, маркетинге (склонность к покупке) и задачах классификации (скоринг)

#### 40. Пробит-модель. Этапы оценки. Области применения.

Пробит-модель – модель бинарного выбора, которая удовлетворяет двум условиям:

1. Остатки модели бинарного выбора  $\varepsilon_i$  являются случайными нормально распределёнными величинами;
2. Функция распределения вероятностей является нормальной вероятностной функцией.

$$p = F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^Z e^{-\frac{u^2}{2}} du,$$

Этапы:

Определение зависимой переменной и факторов

Построение переменной  $Z$ , как линейной комбинации независимых переменных

Построение уравнения для искомой вероятности события и нахождение производных (для оценки кумулятивного и предельного воздействия факторов)

Проведение вычислений с помощью программы (используется метод максимального правдоподобия)

Интерпретация результатов

Качество оценивания

Применяется в следующих областях: медицина, социология, маркетинг, любые статистические исследования.

#### 41. Метод максимального правдоподобия

Метод оценки параметров, который основывается на предположении о том, что вся информация в выборке содержится в функции максимального правдоподобия. Для оценки неизвестного параметра максимизирует функцию максимального правдоподобия. Фактически метод заключается в том, чтобы найти наиболее близкие к реальным данным значения параметров модели. Максимально правдоподобный может использоваться для оценки параметров как линейных, так и нелинейных моделей. В регрессии используется следующая функция:

$$\prod_{i=1}^n [F(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})]^{Y_i} [1 - F(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})]^{1-Y_i}$$

В некоторых случаях для упрощения процесса используют логарифмическую функцию правдоподобия.

$$l(\beta) = \sum_{i=1}^n [Y_i \ln F(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}) + (1 - Y_i) \ln (1 - F(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}))]$$

## 42. Свойства оценок метода максимального правдоподобия

Состоятельные – с увеличением объема наблюдений разница между оценкой и фактическим значением параметров приближаются к нулю

Инвариантные – для полученной оценки  $\theta_L$  параметра  $\theta$  и непрерывной функции  $q(\theta)$ , оценкой значения этой функции будет  $q(\theta_L)$ . Если с помощью метода макс. правдоподобия (ММП) была оценена величина дисперсии показателя ( $\sigma_L^2$ ), то корень из полученной оценки будет оценкой среднего квадратического отклонения, полученной по ММП

асимптотически эффективны - имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра

асимптотически нормально распределены

Могут быть смещенными, то есть точечная оценка, математическое ожидание которой не равно оцениваемому параметру

#### 43. Информационная матрица и оценки стандартных ошибок для оценок параметров logit и probit моделей. Интерпретация коэффициентов в моделях бинарного выбора.

Можно использовать стандартную ошибку оценок, полученных методом максимального правдоподобия. Чтобы оценить параметр требуется реализовать итерационную вычислительную процедуру поиска такого значения оценки параметра, полученного обобщенным методом наименьших квадратов. Задача оценки параметров логит и пробит моделях сводится к анализу линейной функции регрессии.

$$\tilde{y}_j \approx \tilde{\theta}^T \hat{X}_j + \tilde{\varepsilon}_j$$

$\tilde{y}_j = F^{-1}(\hat{p}_j)$  - квантиль уровня  $\hat{p}_j$  функции распределения

В логит и пробит моделях коэффициенты соответствуют предельному (маржинальному) эффекту k- независимой переменной. Этот эффект является функцией всех объясняющих переменных:

$$\frac{\partial \Phi(X_i^T \beta)}{\partial X_{ik}} = \phi(X_i^T \beta) \beta_k; \quad \frac{\partial \Lambda(X_i^T \beta)}{\partial X_{ik}} = \frac{e^{X_i^T \beta}}{(1 + e^{X_i^T \beta})^2} \beta_k$$

Однако знак предельного эффекта j-ой переменной соответствует знаку коэффициента  $\beta_j$  и легко интерпретируется.

#### 44. Мера качества аппроксимации и качества прогноза logit и probit моделей.

Качество построенной логит-регрессии или пробит-регрессии характеризуется с помощью псевдокоэффициента детерминации, который рассчитывается по формуле:

$$\text{psevdoR}^2 = 1 - \frac{1}{\frac{(1 + 2(l_1 - l_0))}{N}}$$

Если значение данного коэффициента близко к единице, то модель регрессии считается адекватной реальным данным.

Мерой качества аппроксимации служат MAPE и MSE.

Средняя абсолютная процентная ошибка (англ. Mean Absolute Percentage Error, MAPE). Это коэффициент, не имеющий размерности, с очень простой интерпретацией. Его можно измерять в долях или процентах.

Если у вас получилось, например, что MAPE=11,4%, то это говорит о том, что ошибка составила 11,4% от фактических значений. Основная проблема данной ошибки — нестабильность.

$$\text{MAPE} = 100\% \times \frac{1}{n} \sum_{i=1}^n \frac{|y_i - a(x_i)|}{|y_i|}$$

MSE применяется в ситуациях, когда нам надо подчеркнуть большие ошибки и выбрать модель, которая дает меньше больших ошибок прогноза. Грубые ошибки становятся заметнее за счет того, что ошибку прогноза мы возводим в квадрат. И модель, которая дает нам меньшее значение среднеквадратической ошибки, можно сказать, что у этой модели меньше грубых ошибок.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (a(x_i) - y_i)^2$$

Также должна иметься BLUE оценка – наилучшая линейная несмещенная оценка, то есть она должна быть несмещенной (мат ожидание остатков =0), эффективной (наименьшая дисперсия отклонений) и состоятельной (точность оценки увеличивается при увеличении размера выборки). Чаще всего меры аппроксимации строятся путем прямого или косвенного сравнения текущей модели и тривиальной модели. Можно выделить и индекс отношения правдоподобия

#### 45. Временные ряды: определение, классификация, цель и задача моделирования временного ряда.

Временной ряд - упорядоченная во времени последовательность численных наблюдений показателя  $y_t$ , где  $t = 1, 2, \dots, n$ , характеризующих уровни развития изучаемого процесса в последовательные моменты времени.

Временные виды классифицируются на два вида: изолированный одномерный ряд – один временной ряд и система несколько взаимосвязанных многомерных рядов.

Целью исследования временного ряда является определение закономерностей в изменении уровней ряда (в динамике) и построение модели изменения с целью дальнейшего прогнозирования.

Задача моделирования временного ряда - разложение наблюдаемых изменений на две составляющие: детерминированную (объясненную) и случайную.

#### 46. Исследование структуры одномерного временного ряда.

Общий случай каждого уровня временного можно представить собой функцию четырех компонент:

$f(t)$ ,  $S(t)$ ,  $C(t)$  и  $\varepsilon(t)$ ,

отражающих как закономерности, так и случайности развития.

$f(t)$  – тренд или тенденция (наиболее важная); плавно меняющаяся компонента, описывающая чистое влияние долговременных факторов, т. е. длительную («вековую») тенденцию изменения признака (например, рост населения, экономическое развитие, изменение структуры потребления и т. п.);

$S(t)$  – сезонная (периодическая) компонента ограниченность времени производства работ в течение года, обусловленную влиянием природных факторов; отражает повторяемость экономических процессов в течение не очень длительного периода (года, иногда месяца, недели и т. д., например объем продаж товаров или перевозок пассажиров в различные времена года);

$C(t)$  – циклическая компонента, колебания с периодом несколько лет; отражает повторяемость экономических процессов в течение длительных периодов (например, влияние волн экономической активности Кондратьева, демографических ям, циклов солнечной активности и т. п.);

$\varepsilon(t)$  – остаточная компонента, отражающая влияние не поддающихся учету и регистрации

Рассчитав несколько коэффициентов автокорреляции, можно определить лаг  $L$ , при котором коэффициент автокорреляции  $r(L)$  наиболее высокий, выявив тем самым структуру временного ряда.

Если наиболее высоким оказывается значение  $r(1)$ , то исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался  $r(L)$ , то ряд содержит (помимо тенденции) колебания с периодом  $L$ . Если ни один из коэффициентов  $r(1)$ ,  $1 = 1, \dots, L$ , не является значимым, можно сделать одно из двух предположений:

1) либо ряд не содержит тенденции и циклических колебаний, а его уровень определяется только случайной компонентой;

2) либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужен дополнительный анализ.

## 47 Функциональные зависимости временного ряда.

### Предварительный анализ временных рядов 47

Чаще всего на практике приходится иметь дело с рядами, включающими три компоненты – тренд, сезонную и случайную составляющие. Такие процессы принято называть тренд-сезонными.

В зависимости от вида связи между компонентами может быть построена: аддитивная модель:  $yt = ft + St + \varepsilon t$  (сумма компонент)

мультипликативная модель временного ряда:  $yt = ft \cdot St \cdot \varepsilon t$  (перемножение компонент)

модель смешанного типа:  $yt = ft \cdot St + \varepsilon t$

### Предварительный анализ

В ходе предварительного анализа динамика исследуемого процесса представляется в графическом формате. При графическом отображении динамики по оси абсцисс откладываются значения переменной  $t$ , а по оси ординат - соответствующие значения показателя  $Y(t)$ .

К процедурам предварительного анализа относятся:

\*выявление аномальных наблюдений.

Аномальные значения временного ряда не отвечают потенциалу исследуемой экономической системы, и их использование для построения трендовой модели может сильно исказить получаемые результаты.

Причинами появления аномальных уровней могут быть технические ошибки при сборе, обработке и передаче информации. Такие ошибки называются ошибками первого рода, их можно выявить и устранить или принять меры к их недопущению. Кроме того, аномальные уровни могут возникать из-за воздействия факторов, имеющих объективный характер, но действующих эпизодически. Такие ошибки называются ошибками второго рода, их невозможно устранить, но можно исключить из рассмотрения, заменив аномальное значение на среднеарифметическое двух соседних уровней.

Аномальные значения можно найти с помощью “Ящика с усами”, метода Ирвинга или метода на основе распределения Стьюдента.

\*проверка наличия тренда. Критерий, основанный на медиане, метод проверки разности средних уровней, метод Фостера-Старта.

\*сглаживание временных рядов. Очень часто уровни экономического ряда динамики колеблются, так что тенденция развития экономического процесса скрыта случайными отклонениями. Сглаживание временного ряда позволяет отфильтровать мелкие случайные колебания и выявить основную тенденцию изменения исследуемой величины. При механическом сглаживании выравнивание отдельных уровней производится с использованием значений соседних уровней.

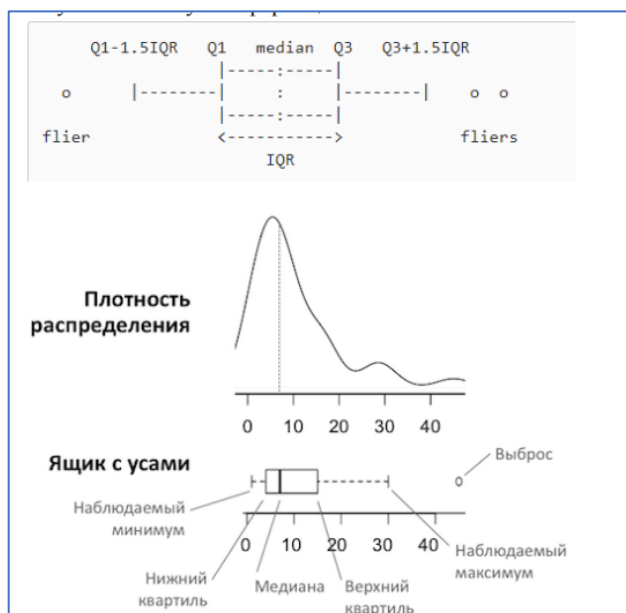
Среднеарифметическое сглаживание, средневзвешанное, среднехронологическое, экспоненциальное.

## 48 Процедура выявления аномальных наблюдений. Причины аномальных значений. Блочные диаграммы по типу «ящика с усами» 48

Предварительная обработка временных рядов состоит в выявлении аномальных значений ряда и сглаживании ряда. Аномальные значения временного ряда не отвечают потенциалу исследуемой экономической системы, и их использование для построения трендовой модели может сильно исказить получаемые результаты. Причинами появления аномальных уровней могут быть технические ошибки при сборе, обработке и передаче информации. Такие ошибки называются ошибками первого рода, их можно выявить и устранить или принять меры к их недопущению. Кроме того, аномальные уровни могут возникать из-за воздействия факторов, имеющих объективный характер, но действующих эпизодически. Такие ошибки называются ошибками второго рода, их невозможно устранить, но можно исключить из рассмотрения, заменив аномальное значение на среднее арифметическое двух соседних уровней.

### Причины аномалий

1. Значения, отражающие объективное развитие процесса, но сильно отличающиеся от общей тенденции, так как они проявляют свои экстремальные воздействия крайне редко. Аномальные значения этой группы не всегда должны исключаться из временного ряда и могут даже оказаться полезными на этапе исследования причинноследственного механизма развития явления. Наличие нехарактерных пиковых значений для одного и того же момента времени в различных временных рядах свидетельствует, как правило, о причинных связях между соответствующими показателями.
2. Значения, возникающие вследствие изменений методики расчета. Нехарактерные значения второй группы не должны исключаться из рассмотрения, а приниматься за «повторные» (пороговые), начиная с которых должны быть пересчитаны по новой методике все предыдущие значения временного ряда.
3. Значения, возникающие вследствие ошибок при измерении показателя, при записи и передаче информации, а также значения, связанные с различными катастрофическими явлениями, не влияющими на дальнейший ход развития явления. Аномальные значения третьей группы должны быть исключены из рассмотрения в любом случае, так как они искажают представление о характере развития явления и могут оказать существенное влияние на выводы, полученные в результате анализа ряда, содержащего такую искаженную информацию.



Границами ящика служат первый и третий квартили (25-й и 75-й процентиля соответственно), линия в середине ящика — медиана (50-й процентиль). Концы усов — края статистически значимой выборки (без выбросов), и они могут определяться несколькими способами. Наиболее распространённые значения, определяющие длину «усов»:

- Минимальное и максимальное наблюдаемые значения данных по выборке (в этом случае выбросы отсутствуют);
  - Разность первого квартиля и полутора межквартильных расстояний; сумма третьего квартиля и полутора (1.5) межквартильных расстояний (в этом случае присутствуют выбросы). В общем виде эта формула имеет вид
- Данные, выходящие за границы усов (выбросы), отображаются на графике в виде точек,

маленьких кружков или звёздочек.

Особенности. Хорошо подходит для небольших массивов данных. Показывает экстремальные значения (аномалия характеризуется, как правило, не только экстремальными значениями отдельных признаков). Требуется дополнительный анализ

## 49 Процедура выявления аномальных наблюдений на основе распределения Стьюдента. Особенности применения метода. Анализ аномальных наблюдений 49

По имеющейся выборке вычисляется статистика:

$$\hat{\tau} = \frac{|x^* - \bar{x}|}{\bar{S}},$$

где  $x^* = \text{Arg max}_i |x_i - \bar{x}|$  — наблюдение, предположительно являющееся аномальным;  $\bar{x}$ ,  $\bar{S}$  — среднее и среднеквадратическое отклонение соответственно.

Величина  $\tau$  называется максимальным относительным отклонением. Далее проводится сравнение  $\tau$  с критическими значениями

$$\tau_{\alpha,n} = \frac{t_{\text{кр}}\left(1 - \frac{\alpha}{2}, n - 2\right) \sqrt{n - 1}}{\sqrt{n - 2 + t_{\text{кр}}^2\left(1 - \frac{\alpha}{2}, n - 2\right)}}$$

где  $t_{\text{кр}}\left(1 - \frac{\alpha}{2}, n - 2\right)$  — критическое значение распределения Стьюдента при различных значениях уровня значимости  $\alpha$  и количеством степеней свободы  $n - 2$ .

Максимальные относительные отклонения в процессе вычисления разделяют на три группы в соответствии с условиями:

**Группа 1:**

$$\hat{\tau} \leq \tau_{0,05,n}$$

наблюдение нельзя считать аномальным.

**Группа 2:**

$$\tau_{0,05,n} < \hat{\tau} < \tau_{0,001,n}$$

наблюдение может быть признано аномальным, если в пользу этого имеются и другие поводы (например, на основе изучения аналогов и др. методы).

**Группа 3:**

$$\hat{\tau} > \tau_{0,001,n}$$

наблюдение признается аномальным.

Обнаруженные аномалии стоит изучить: посмотреть, вызваны ли они каким-то событием, системной ошибкой или чем-то другим. Понимая причину, значение стоит заменить (на медиану, среднее арифметическое по соседям или на что-то другое, подходящее по смыслу) или попробовать восстановить, если это опечатка.

**Особенности.** Выборка небольшого объема  $n \leq 25$ . Если попало во вторую группу, можно случайно сделать неправильный вывод, требуются дополнительные тесты. Если наблюдается стабильный рост, тогда начальные и конечные наблюдения будут выбиваться, так как Стьюдент смотрит на середину ряда. Если 2020 год - аномалия, то может показать, что 2019 год - тоже аномалия. Стоит использовать, когда имеем дело с ровным, небыстро растущим трендом.

При моделировании временного ряда часто отбрасываются аномальные наблюдения, резко отклоняющиеся от направления эволюции ряда. Такого рода выбросы, вместо исключения, можно моделировать с помощью фиктивных переменных, соответствующих фиксированным моментам времени.

Предположим, что в момент  $t^*$  в экономике произошло какое-нибудь важное событие (например, отставка правительства). Тогда можно построить фиктивную переменную  $\delta_t^{t^*}$ , которая равна нулю всегда, кроме момента  $t = t^*$ , когда она равна единице:  $\delta_t^{t^*} = (0, \dots, 0, 1, 0, \dots, 0)$ .

Такая фиктивная переменная пригодна только для моделирования кратковременного отклонения временного ряда. Если же в экономике произошел структурный сдвиг, вызвавший скачок в динамике ряда, то следует использовать фиктивную переменную другого вида:  $(0, \dots, 0, 1, \dots, 1)$ . Эта переменная равна нулю до некоторого фиксированного момента  $t^*$ , а после этого момента становится равной единице.

Заметим, что последние два вида переменных нельзя использовать для прогнозирования, поскольку они относятся к единичным непрогнозируемым событиям.

## 50 Процедура выявления аномальных наблюдений на основе метода Ирвина. Особенности применения метода. Анализ аномальных наблюдений

Для всех или только для подозреваемых в аномальности наблюдений вычисляется величина  $\lambda_t$ :

$$\lambda_t = \frac{|y_t - y_{t-1}|}{S_y}$$

где  $S_y = \sqrt{\frac{\sum_{t=1}^n (y_t - \bar{y})^2}{n-1}}$      $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$

Если рассчитанная величина  $\lambda_t$ , превышает табличный уровень, то уровень  $y_t$ , считается аномальным. Аномальные наблюдения необходимо исключить из временного ряда и заменить их расчетными значениями.

Обнаруженные аномалии стоит изучить: посмотреть, вызваны ли они каким-то событием, системной ошибкой или чем-то другим. Понимая причину, значение стоит заменить (на медиану, среднее арифметическое по соседям или на что-то другое, подходящее по смыслу) или попробовать восстановить, если это опечатка.

Особенности. Табличные значения критерия Ирвина рассчитаны не для всех объемов выборки (но есть методы аппроксимации). Ирвин смотрит по отношению с предыдущим значением, поэтому если 2020 -аномалия, то он “потянет” аномальность на 2021 год. Ирвин не подходит для сезонных данных, так как идет вычитание из предыдущего.

При моделировании временного ряда часто отбрасываются аномальные наблюдения, резко отклоняющиеся от направления эволюции ряда. Такого рода выбросы, вместо исключения, можно моделировать с помощью фиктивных переменных, соответствующих фиксированным моментам времени.

Предположим, что в момент  $t^*$  в экономике произошло какое-нибудь важное событие (например, отставка правительства). Тогда можно построить фиктивную переменную  $\delta_t^{t^*}$ , которая равна нулю всегда, кроме момента  $t = t^*$ , когда она равна единице:  $\delta_t^{t^*} = (0, \dots, 0, 1, 0, \dots, 0)$ .

Такая фиктивная переменная пригодна только для моделирования кратковременного отклонения временного ряда. Если же в экономике произошел структурный сдвиг, вызвавший скачок в динамике ряда, то следует использовать фиктивную переменную другого вида:  $(0, \dots, 0, 1, \dots, 1)$ . Эта переменная равна нулю до некоторого фиксированного момента  $t^*$ , а после этого момента становится равной единице.

Заметим, что последние два вида переменных нельзя использовать для прогнозирования, поскольку они относятся к единичным непрогнозируемым событиям.

## 51 Проверка наличия тренда. Критерий серий, основанный на медиане. Особенности применения метода 51

Тренд, или основная тенденция.  $f(t)$  - это устойчивая закономерность, наблюдаемая в течение длительного периода времени. (Длительная, “вековая” закономерность изменения уровней). Тренд пределяет общее направление развития экономического процесса.

Этап 1:

Элементы исходного ряда располагаются в порядке возрастания, т.е. из исходного ряда  $y_t$  длиной  $n$  образуется ранжированный (вариационный)  $y_t$

Этап 2:

Определяется медиана ранжированного ряда

$$Me = \begin{cases} y_{m+1}, & n = 2m + 1 - \text{нечетное} \\ \frac{(y_m + y_{m+1})}{2}, & n = 2m - \text{четное} \end{cases}$$

Этап 3:

Образуется последовательность  $\delta_i$  из плюсов и минусов

$$\delta_i = \begin{cases} +, & \text{если } y_t > Me, \quad t = 1, \dots, n \\ -, & \text{если } y_t < Me, \quad t = 1, \dots, n \end{cases}$$

Если значение  $y_t$  равно медиане, то это значение пропускается.

Этап 4:

Подсчитывается:

$v(n)$  – число серий в совокупности  $\delta_i$ , где под серией понимается последовательность подряд идущих плюсов или минусов. Один плюс или один минус тоже будет считаться серией.  $\tau_{\max}(n)$  – протяженность самой длинной серии.

Этап 5:

Проверка гипотезы: при условии случайности ряда (при отсутствии систематической составляющей) протяженность самой длинной серии не должна

быть слишком большой, а общее число серий - слишком маленьким.

Чтобы не была отвергнута гипотеза о случайности исходного ряда (об отсутствии

систематической составляющей), должны выполняться следующие неравенства.

(для 5% уровня значимости квантиль нормального распределения равен

$a_{кр} =$

1,96):

$$\begin{cases} \tau_{\max}(n) < [3.3(\lg n + 1)] \\ v(n) > [\frac{1}{2}(n + 1 - 1.96\sqrt{(n - 1)})] \end{cases}$$

Критерий серий, основанный на медиане, улавливает только монотонное изменение среднего (оценки математического ожидания).

## 52 Проверка наличия тренда. Метод проверки разности средних уровней. Особенности применения метода 52

Тренд, или основная тенденция.  $f(t)$  - это устойчивая закономерность, наблюдаемая в течение длительного периода времени. (Длительная, “вековая” закономерность изменения уровней). Тренд определяет общее направление развития экономического процесса.

Этап 1:

Исходный временной ряд  $Y_1, Y_2, Y_3, \dots, Y_n$  разбивается на две примерно равные по числу уровней части: в первой части  $n_1$  первых уровней исходного ряда, во второй  $n_2$  остальных уровней ( $n_1 + n_2 = n$ ).

Этап 2:

Для каждой из этих частей вычисляются средние значения и дисперсии:

$y^*$  ( $y_{cp}$ )

$$y_1^* = \frac{\sum_{t=1}^{n_1} y_t}{n_1} \quad \sigma_1^2 = \frac{\sum_{t=1}^{n_1} (y_t - y_1^*)^2}{n_1 - 1}$$
$$y_2^* = \frac{\sum_{t=n_1+1}^{n_2} y_t}{n_2} \quad \sigma_2^2 = \frac{\sum_{t=n_1+1}^{n_2} (y_t - y_2^*)^2}{n_2 - 1}$$

Этап 3

Проверка равенства (однородности) дисперсий обеих частей ряда с помощью F-критерия Фишера, которая основана на сравнении расчетного значения этого критерия:

$$F = \begin{cases} \frac{\sigma_1^2}{\sigma_2^2}, & \text{если } \sigma_1^2 > \sigma_2^2 \\ \frac{\sigma_2^2}{\sigma_1^2}, & \text{если } \sigma_1^2 < \sigma_2^2 \end{cases}$$

с табличным (критическим) значением критерия Фишера  $F_{кр}$  с заданным уровнем значимости (уровнем ошибки)  $\alpha$ . В качестве,  $\alpha$  чаще всего берут значения ОД (10%-ная ошибка), 0,05 (5%-ная ошибка), 0,01 (1%-ная ошибка). Величина

$(1-\alpha)$  называется доверительной вероятностью. Если расчетное значение  $F$  меньше табличного  $F_{кр}$ , то гипотеза о равенстве дисперсий принимается и переходят к четвертому этапу. Если  $F$  больше или равно  $F_{кр}$ , гипотеза о равенстве дисперсий отклоняется и делается вывод, что данный метод для определения наличия тренда ответа не дает.

Этап 4:

Проверяется гипотеза об отсутствии тренда с использованием  $t$ -критерия Стьюдента. Для этого определяется расчетное значение критерия Стьюдента по формуле:

$$t = \frac{|\bar{y}_1 - \bar{y}_2|}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

где  $\sigma$  – среднеквадратическое отклонение разности средних:

$$\sigma = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

Если расчетное значение  $t$  меньше табличного значения статистики (Стьюдента)  $t_{кр}$  с заданным уровнем значимости,  $\alpha$ , гипотеза принимается, т.е. тренда нет, в противном случае тренд есть. Заметим, что в данном случае табличное значение  $t_{кр}$  берется для числа степеней свободы, равного  $n_1 + n_2 - 2$ , при этом данный метод применим только для рядов с монотонной тенденцией.

Применим только для рядов с **монотонной** тенденцией. Если же ряд меняет общее направление развития, то точка поворота тенденции может оказаться близкой к середине ряда, в силу этого средние двух отрезков ряда будут близки и проверка может не показать наличие тренда

Если расчетное значение  $F$  меньше табличного  $F_{кр}$ , то гипотеза о равенстве дисперсий принимается и переходят к четвертому этапу. Если  $F$  больше или равно  $F_{кр}$ , гипотеза о равенстве дисперсий отклоняется и делается вывод, что данный метод для определения наличия тренда **ответа не дает**.

## 53 Проверка наличия тренда. Метод Фостера Стьюарта.

### Особенности применения метода

Тренд, или основная тенденция.  $f(t)$  - это устойчивая закономерность, наблюдаемая в течение длительного периода времени. (Длительная, “вековая” закономерность изменения уровней). Тренд определяет общее направление развития экономического процесса.

Кроме самого тренда он позволяет установить наличие тренда дисперсии.

При отсутствии тренда дисперсии разброс уровней ряда постоянен.  
при наличии тренда дисперсии дисперсия увеличивается или уменьшается.  
Определим две последовательности:

$$k_t = \begin{cases} 1, & \text{если } y_t \text{ больше всех предыдущих уровней} \\ 0, & \text{в противном случае} \end{cases}$$
$$l_t = \begin{cases} 1, & \text{если } y_t \text{ меньше всех предыдущих уровней} \\ 0, & \text{в противном случае} \end{cases}$$

Вычислим величины  $s$  и  $d$ , характеризующие изменение временного ряда и дисперсии:

$$s = \sum_{t=2}^n (k_t + l_t) \quad d = \sum_{t=2}^n (k_t - l_t)$$

Величина  $s$  характеризует изменение временного ряда, она может принимать значение от 0 (когда все уровни ряда равны) до  $n - 1$  (ряд монотонный).

Величина  $d$  характеризует изменение дисперсии временного ряда и изменяется от  $-(n-1)$  (когда ряд монотонно убывает) до  $(n-1)$  (когда ряд монотонно возрастает).

Эти величины являются случайными с математическим ожиданием  $\mu$  для значения  $s$  и 0 для значения  $d$ .

Проверим гипотезы о случайности отклонения величин  $s$  от её математического ожидания  $\mu$  и о случайности отклонения величины  $d$  от нуля с помощью критерия Стьюдента для средней и для дисперсии:

$$t_s = \frac{|s - \mu|}{\sigma_1}, \sigma_1 = \sqrt{2 \ln n - 3,4253}$$
$$t_d = \frac{|d - 0|}{\sigma_2}, \sigma_2 = \sqrt{2 \ln n - 0,8456}$$

Где  $\mu$  - математическое ожидание величины  $S$  для случайного временного ряда;

$\sigma_1$  – среднее квадратичное отклонение  $s$  для случайного временного ряда;

$\sigma_2$  – среднее квадратичное отклонение  $d$  для случайного временного ряда;

Если  $t_{кр}$  больше расчетного значения  $t_{кр} > t_s, t_{кр} > t_d$ , то соответствующий тренд отсутствует.

## 54 Сглаживание временных рядов. Простая (среднеарифметическая) скользящая средняя. Взвешенная (средневзвешенная) скользящая средняя. Среднехронологическая. Экспоненциальное сглаживание 54

Очень часто уровни экономического ряда динамики колеблются, так что тенденция развития экономического процесса скрыта случайными отклонениями. Сглаживание временного ряда позволяет отфильтровать мелкие случайные колебания и выявить основную тенденцию изменения исследуемой величины. При механическом сглаживании выравнивание отдельных уровней производится с использованием значений соседних уровней. Если есть сильно выраженная сезонность, сглаживать не имеет смысла, так как прогноз тогда будет неверным.

Для сглаживания используются следующие методы:

- \* Простая (среднеарифметическая) скользящая средняя
- \* Взвешенная (средневзвешенная) скользящая средняя
- \* Среднехронологическая
- \* Экспоненциальное сглаживание

Простая (среднеарифметическая) скользящая средняя

$$\tilde{y}_t = \frac{\sum_{i=t-p}^{t+p} y_i}{2p+1}, p < t < n-p$$

Сглаженное значение  $y_t$  является среднеарифметическим из  $2p+1$  соседних точек. Наиболее часто используется сглаживание по 5 точкам: DD

$$\tilde{y}_t = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5}$$

Взвешенная (средневзвешенная) скользящая средняя

$$\tilde{y}_t = \frac{\sum_{i=t-p}^{t+p} p_i y_i}{\sum_{i=t-p}^{t+p} p_i}, p < t < n-p$$

В этом методе каждая из точек входит в общую сумму с весовым коэффициентом  $p_i$ .

Для сглаживания по 5 точка используют весовые коэффициенты (-3, 12, 17, 12, -3).

Для сглаживания по 7 точка используют коэффициенты (-2, 3, 6, 7, 6, 3, -2) или (5, -30, 75, 131, 75, -30, 5)

Среднехронологическая

$$\tilde{y}_t = \frac{\frac{y_t - T/2}{2} + \sum_{i=t-p}^{t+p} y_i + \frac{y_t + T/2}{2}}{T}, \frac{T}{2} < t < n - \frac{T}{2}$$

Эта формула используется для моментных временных рядов. Обычно период сглаживания принимают равным году, т.е.  $T = 4$  квартала или  $T = 12$  месяцев.

Экспоненциальное сглаживание 1

Используются все предшествующие точки, причем значения весовых коэффициентов убывают по экспоненте по мере удаления от текущей точки. Текущая точка зависит от всех предыдущих точек:

$$\tilde{y}_t = \frac{\sum_{i=1}^t p_i y_i}{\sum_{i=1}^t p_i}$$

В таком виде она неудобна для использования, поскольку для каждой точки необходим свой набор весовых коэффициентов. Используя рекуррентные соотношения, получим выражение для текущей сглаженной точки как функцию от текущей несглаженной точки и предыдущей сглаженной:

$$\tilde{y}_t = \alpha y_t + (1 - \alpha) \tilde{y}_{t-1}, \quad 0 < \alpha < 1$$

Где  $\alpha$  – параметр сглаживания;  $(1-\alpha)$  – коэффициент дисконтирования.

Фиктивное начальное значение сглаженного ряда принимают равным первой точке или среднеарифметическому первых трех точек:

$$\tilde{y}_0 = y_1 \text{ или } \tilde{y}_0 = (y_1 + y_2 + y_3)/3$$

При сглаживании временного ряда по  $2p+1$  соседним точкам  $p$  в начале и в конце ряда остаются несглаженными.

Эти точки следует либо исключить из рассмотрения, либо использовать для них специальные формулы сглаживания для крайних точек. В частности, для сглаживания по трем точкам можно использовать формулы:

$$\tilde{y}_1 = \frac{(5y_1 + 2y_2 - y_3)}{6}; \quad \tilde{y}_n = \frac{(5y_n + 2y_{n-1} - y_{n-2})}{6}$$

Заметим, что при экспоненциальном сглаживании не теряются ни начальные, ни конечные точки

## 55. Трендовые модели. Без предела роста. Примеры функций. Содержательная интерпретация параметров.

Трендовые модели используются для моделирования динамических процессов, структура которых включает две составляющие: детерминированную, которую возможно описать с помощью гладкой аналитической функции, и случайную компоненту. (из файла Методы и модели) На практике для описания тенденции развития явления широко используются модели кривых роста, представляющие собой различные функции времени. При таком подходе изменение исследуемого показателя связывают лишь с течением времени; считается, что влияние других факторов несущественно или косвенно сказывается через фактор времени. Правильно выбранная модель кривой роста должна соответствовать характеру изменения тенденции исследуемого явления. Кривая роста позволяет получить выровненные или теоретические значения уровней динамического ряда. Это те уровни, которые наблюдались бы в случае полного совпадения динамики явления с кривой. Прогнозирование на основе модели кривой роста базируется на экстраполяции, т.е. на продлении в будущее тенденции, наблюдавшейся в прошлом. При этом предполагается, что во временном ряду присутствует тренд, характер развития показателя обладает свойством инерционности, сложившаяся тенденция не должна претерпевать существенных изменений в течение периода упреждения. (конец копипаста из файла) В этом случае легко получить оценки будущих состояний уровней ряда используя метод экстраполяции. Плавную кривую (гладкую функцию), аппроксимирующую временной ряд принято называть *кривой роста*.

На практике используются кривые роста, которые позволяют описывать процессы трех основных типов:

\* без предела роста;

\* с пределом роста без точки перегиба; \* с пределом роста и точкой перегиба.

Без предела роста. Функции, используемые для описания процессов с монотонным характером тенденции развития и отсутствием пределов роста. Эти условия справедливы для многих экономических показателей, например, для большинства натуральных показателей промышленного производства.

Примеры функций процессов без предела роста:

\* прямая (полином первой степени)  $y_t = a_0 + a_1 t$ , - используется для описания процессов, развивающихся во времени равномерно;

\* парабола (полином второй степени)  $y_t = a_0 + a_1 t + a_2 t^2$ , - применим в тех случаях, когда процесс

развивается равноускоренно (т.е. имеется равноускоренный рост или равноускоренное снижение уровней);

\* экспонента  $y_t = \exp(a_0 + a_1 t)$ .

Обычно в экономических исследованиях применяются полиномы не выше третьего порядка. Использовать для определения тренда полиномы высоких степеней нецелесообразно, поскольку полученные таким образом аппроксимирующие функции будут отражать случайные отклонения (что противоречит смыслу тенденции).

Кривую роста также можно использовать для восстановления пропущенных значений, и тогда полином может быть любой степени.

---

Содержательная интерпретация параметров:

---

\*  $a_0$  во всех моделях без предела роста задает начальные условия развития, а в моделях с пределом роста - асимптоту функций,

\*  $a_1$  определяет скорость или интенсивность развития,

\*  $a_2$  - изменение скорости или интенсивности развития.

Оценки параметров полиномов выполняют методом наименьших квадратов. При использовании кривых роста, не являющихся полиномами, необходимо применять замену переменных, позволяющую определить параметры линии роста с помощью системы линейных уравнений

## 56. Трендовые модели. С пределом роста без точки перегиба. Примеры функций. Содержательная интерпретация параметров.

Трендовые модели используются для моделирования динамических процессов, структура которых включает две составляющие: детерминированную, которую возможно описать с помощью гладкой аналитической функции, и случайную компоненту. В этом случае легко получить оценки будущих состояний уровней ряда используя метод экстраполяции.

Модифицированная экспонента относится к кривым *с пределом роста без точки перегиба* и имеет вид:  $f(t) = k + ab^t$ .

где  $a < 0$ ;  $0 < b < 1$ ;  $k$  – асимптота, значение которой считается известным.

Параметры  $a$  и  $b$  можно найти, как и для простой экспоненты, перенеся  $k$  в левую часть:  $f(t) - k = ab^t$ .

Более сложным вариантом экспоненциальной кривой является логарифмическая парабола:  $f(t) = a_0 a_1^t a_2^{t^2}$ .

Прологарифмировав это выражение, получим параболу:

$$\ln y_t = \ln a_0 + t \ln a_1 + t^2 \ln a_2.$$

Таким образом, оценку параметров логарифмической параболы можно опять осуществить с помощью метода наименьших квадратов, используя систему нормальных уравнений для параболы.

## 57. Трендовые модели. С пределом роста и точкой перегиба или кривые насыщения. Примеры функций. Содержательная интерпретация параметров.

Трендовые модели используются для моделирования динамических процессов, структура которых включает две составляющие: детерминированную, которую возможно описать с помощью гладкой аналитической функции, и случайную компоненту. В этом случае легко получить оценки будущих состояний уровней ряда используя метод экстраполяции.

К кривым *с пределом роста и точкой перегиба* относятся кривая Гомперца и логистическая кривая (Перла-Рида). Кривая Гомперца имеет вид:

$$f(t) = ka^{b^t}.$$

Если  $a > 1$ , асимптота, равная  $k$ , лежит ниже кривой, а сама кривая изменяется монотонно: при  $b < 1$  – монотонно убывает; при  $b > 1$  – монотонно возрастает.

В кривой Гомперца выделяют четыре участка: на первом – прирост функции незначителен, на втором – прирост увеличивается, на третьем участке прирост примерно постоянен, на четвертом – происходит замедление темпов прироста и функция приближается к значению  $k$ .

Применяя дважды логарифмирование, получим линейное уравнение:

$$\ln f(t) = \ln k + b^t \ln a, \quad \ln \ln(\ln f(t) - \ln k) - \ln \ln a + t \ln b.$$

Уравнение логистической кривой получается путем замены в модифицированной экспоненте  $f(t)$  обратной величиной  $\frac{1}{f(t)}$ :

$$\frac{1}{f(t)} = k + ab^t.$$

Используются и другие формы записи уравнения логистической кривой:

$$f(t) = \frac{k}{1 - ae^{-bt}}, \quad f(t) = \frac{k}{1 + ab^t}, \quad f(t) = \frac{k}{1 + 10^{a-bt}}.$$

При  $t \rightarrow -\infty$  логистическая кривая стремится к нулю, а при  $t \rightarrow \infty$  – к асимптоте, равной значению параметра  $k$ . Кривая симметрична относительно точки перегиба с координатами  $\frac{t = \ln b}{a}$ ;  $f(t) = \frac{k}{2}$ .

## 58. Выбор кривой роста.

Для выбора кривой роста используется метод характеристик прироста, основанный на использовании отдельных характерных свойств рассмотренных выше кривых. Процедура выбора кривых с использованием этого метода включает выравнивание ряда  $Y_t$  с помощью скользящей средней (обычно среднеарифметической по трем точкам) и определение средних приростов и производных величин:

$\Delta y_t = \frac{y_{t+1} - y_{t-1}}{2}$  – первый средний прирост;

$\Delta^2 y_t = \frac{\Delta y_{t+1} - \Delta y_{t-1}}{2}$  – второй средний прирост;

$\frac{\Delta y_t}{y_t}$ ,  $\ln \Delta y_t$ ,  $\ln \frac{\Delta y_t}{y_t}$ ,  $\ln \frac{\Delta y_t}{y_t^2}$  – производные величины.

В соответствии с характером изменений средних приростов и производных показателей выбирается вид кривой роста с помощью табл. №.

Показатель	Характер изменения	Кривая роста
$\Delta y_t$	Примерно постоянный	Полином первого порядка
$\Delta y_t$	Примерно линейный	Полином второго порядка
$\Delta^2 y_t$	Примерно линейный	Полином третьего порядка
$\frac{\Delta y_t}{y_t}$	Примерно постоянный	Экспонента
$\ln \Delta y_t$	Примерно линейный	Модифицированная экспонента
$\ln \frac{\Delta y_t}{y_t}$	Примерно линейный	Кривая Гомперца
$\ln \frac{\Delta y_t}{y_t^2}$	Примерно линейный	Логистическая кривая

На практике отбирают две-три кривые роста и окончательный вывод делают исходя из значений критерия, в качестве которого принимают сумму квадратов отклонений фактических значений уровней от расчетных. Из рассматриваемых кривых предпочтение будет отдано той, которой соответствует минимальное значение критерия. Это удобно еще и потому, что параметры выбранной кривой роста определяются методом наименьших квадратов.

## 59. Прогнозирование с помощью кривой роста.

**Прогнозирование с помощью кривой роста.** Прогнозирование состоит из двух этапов. Для вычисления точечного прогноза в построенную модель подставляют соответствующие значения фактора  $t = n + k$ , входящие в интервал упреждения. Например, для линейной модели прогноз выглядит следующим образом:

$$\hat{y}_{\text{прогноз}(n+k)} = a_0 + a_1(n + k)$$

Для построения интервального прогноза рассчитывают доверительный интервал. Для этого вначале необходимо определиться с уровнем доверия к будущему прогнозу. Примем, например, уровень значимости  $\alpha = 0,1$ , следовательно, доверительная вероятность равна 90%. Ширину доверительного интервала для линейной трендовой модели определяют так:

$$U(k) = S_{\hat{y}} t_{\alpha} \sqrt{1 + \frac{1}{n} + \frac{(n+k-\bar{t})^2}{\sum_{t=1}^n (t-\bar{t})^2}},$$

где  $S_{\hat{y}} = \sqrt{\frac{\sum_{t=1}^n \varepsilon_t^2}{n-m-1}}$ . Далее вычисляют верхнюю и нижнюю границы прогноза следующим образом:

Верхняя граница =  $y_{\text{прогноз}} + U(k)$

Нижняя граница =  $y_{\text{прогноз}} - U(k)$ .

Если же в качестве трендовой модели выбрана нелинейная модель, которую можно линеаризовать, приведя к линейной многофакторной форме, то ширину доверительного интервала можно определить таким способом:

$$U(k) = S_{\hat{y}} t_{\alpha} \sqrt{1 + \vec{x}_0^T (X^T X)^{-1} \vec{x}_0}$$

где  $\vec{x}_0$  – вектор прогнозных оценок регрессоров.

## 60. Прогнозирование временного ряда на основе трендовой модели.

**При экстраполяционном прогнозировании (более сложное)** экономических процессов необходимо определить два элемента: точечный и интервальный прогнозы.

Точечный прогноз – это значение экономического показателя в будущем, определенное путем подстановки значения времени в уравнение выбранной кривой роста. Совпадение фактических данных в будущем и точечного прогнозного значения маловероятно. Поэтому точечный прогноз дополняют двухсторонними границами, т.е. таким интервалом, в котором с большой степенью вероятности ожидается фактическое значение прогнозируемого показателя. Такой прогноз называется интервальным, он определяется с помощью доверительного интервала

$$\hat{y}_{\text{факт}}(t) = f_t + U(k),$$

где  $\hat{y}_{\text{факт}}(t)$  – фактическое значение в будущем;  $U(k)$ , – доверительный интервал.

Величина доверительного интервала зависит от стандартной ошибки аппроксимации временного ряда с помощью кривой роста, от времени упреждения прогноза, от длины временного ряда и от уровня значимости прогноза.

Стандартная ошибка аппроксимации прогнозируемого показателя определяется выражением

$$S_{\hat{y}} = \sqrt{\frac{\sum_{t=1}^n (y_t - f_t)^2}{n-l}},$$

где  $l$  – число параметров трендовой модели.

Для линейного тренда доверительный интервал определяется формулой

$$U(k) = S_{\hat{y}} t_{\alpha} \sqrt{1 + \frac{1}{n} + \frac{3(n+2k-1)^2}{n(n^2-1)}}$$

где  $k$  – период упреждения, т.е. число шагов, на которые делается прогноз;  $t_{\alpha}$  – критерий Стьюдента для числа степеней свободы  $n-2$  уровня значимости  $\alpha$ .

Для полиномов второго и третьего порядка используется выражение, в котором начало отсчета времени перенесено на середину временного ряда наблюдений:

$$U(k) = S_{\hat{y}} t_{\alpha} \sqrt{1 + \frac{1}{n} + \frac{t_k^2}{\sum_{t=1}^n t^2} + \frac{\sum_{t=1}^n t^4 - 2t_k^2 \sum_{t=1}^n t^2}{n \sum_{t=1}^n t^4 - (\sum_{t=1}^n t^2)^2} \frac{t^2 + nt_k^4}{t^2}},$$

где  $t_k$  – время прогноза, а суммирование выполняется по всем значениям временного ряда

$$-\frac{n-1}{2} < t < \frac{n-1}{2}.$$

Несмотря на то, что приведенные формулы позволяют определить прогноз на любое число шагов, попытка заглянуть слишком далеко приведет к очень большим ошибкам. Длина периода упреждения не должна превышать одной трети длины ряда наблюдений.

## 61. Адаптивная модель прогнозирования Брауна.

Адаптивные модели прогнозирования – это модели, способные приспособливать свою структуру и параметры к изменению свойств моделируемого процесса. Как и в трендовых моделях, основным фактором в адаптивных моделях является время, но наблюдениям (уровням ряда) придаются различные веса в зависимости от силы их влияния на текущий уровень ряда. Это позволяет учитывать изменения в тенденции ряда, а также колебания.

Все адаптивные модели основаны на двух схемах: скользящего среднего и авторегрессии. В моделях скользящего среднего текущий уровень является средневзвешенной суммой всех предыдущих уровней. Такие модели хорошо отражают изменение тенденции, но не позволяют отражать колебания. В авторегрессионных моделях для расчета текущего уровня используются не все, а только несколько последних значений ряда, при этом значения весовых коэффициентов определяются не их близостью к моделируемому уровню, а теснотой связи между уровнями. Наиболее часто для краткосрочного прогнозирования изменяющихся процессов используется адаптивная модель Брауна. Она позволяет отображать развитие линейной или параболической тенденции, а также рядов без тенденции. Соответственно различают модели нулевого (наивная), первого или второго порядков вида:

$$\begin{aligned}y_{t+k} &= A_0, \\y_{t+k} &= A_0 + A_1 k, \\y_{t+k} &= A_0 + A_1 k + A_2 k^2,\end{aligned}$$

где  $t$  – текущее время;  $k$  – время упреждения. Порядок модели определяется априорно из предварительного анализа временного ряда и законов развития прогнозируемого процесса.

Модель первого порядка строится следующим образом.

1. По нескольким первым точкам методом наименьших квадратов найдем значения параметров  $A_0, A_1$  линейной модели (или зададим их):

$$\hat{y}_{\text{прогноз}}(t) = A_0 + A_1 t.$$

2. Используя найденные параметры, найдем прогнозное значение на следующем шаге:

$$\hat{y}_{\text{прогноз}}(t+k) = A_0(t) + A_1(t)k, \quad k = 1.$$

Найдем ошибку прогнозирования:

$$e(t+k) = y(t+k) - \hat{y}_{\text{прогноз}}(t+k).$$

4. В соответствии с ошибкой изменим значения параметров модели:

$$\begin{aligned}A_0(t+1) &= A_0(t) + A_1(t) + (1-\beta)^2 e(t), \\A_1(t+1) &= A_1(t) + (1-\beta)^2 e(t),\end{aligned}$$

где  $\beta$  – коэффициент дисконтирования данных,  $0 < \beta < 1$ .

5. По модели с исправленными параметрами найдем прогноз на следующий шаг и вернемся к п. 3, если  $t < N$  (т.е. время обучения модели еще не завершилось), при  $t \geq N$  будем использовать полученное значение как прогнозное, не изменяя параметров модели.

6. Дополним точечный прогноз интервальным:

$$U(k) = S_{\hat{y}} t_{\alpha} \sqrt{1 + \frac{1}{n} + \frac{3(n+2k-1)^2}{n(n^2-1)}}$$

где  $t_{\alpha}$  – значение критерия Стьюдента;  $S_{\hat{y}}$  – среднее квадратичное отклонение прогнозируемого показателя;  $n$  – число наблюдений ряда.

## 62. Моделирование тренд-сезонных процессов. Типы функциональных зависимостей.

Схема не определяет методов решения каждой задачи, методы могут изменяться, совершенствоваться со временем, но она определяет совокупность и последовательность вопросов, которые должны быть решены для полного исследования сезонного временного ряда.



Упорядоченная во времени последовательность наблюдений экономического процесса называется временным рядом, и если процесс подвержен периодическим колебаниям, имеющим определенный и постоянный период, равный годовому промежутку, то мы имеем дело с тренд-сезонным временным рядом (сезонным временным рядом). Существуют аддитивные и мультипликативные модели динамики, в которой присутствует не только основная тенденция, но и сезонные (периодические) отклонения от нее. Рассмотрим мультипликативную модель, т.к. с ее помощью можно не только выделить наличие сезонных колебаний, но и учесть возможное изменение их амплитуды. Адаптивные методы позволяют дифференцировать информационную ценность уровней временного ряда в зависимости от степени устаревания данных.

### 63. Модель Хольта-Уинтерса (адаптивная модель).

Адаптивными методами прогнозирования называются методы, позволяющие строить самокорректирующиеся экономико-математические модели, которые учитывают результат реализации прогноза, сделанного на предыдущем шаге, а так же различают информационную ценность членов временного ряда, благодаря чему способны реагировать на изменяющиеся условия, и на этой основе давать на ближайшую перспективу более точные прогнозы.

Инструментом прогноза в адаптивных методах прогнозирования служат математические модели, первоначальная оценка параметров которых осуществляется обычно по некоторой выборке исходного ряда, называемого обучающей последовательностью.

Метод Хольта, применительно к сезонным эффектам был развит П. Уинтерсом. Модель Хольта-Уинтерса (или по другому адаптивная сезонная модель с линейным трендом) содержит уже три параметра адаптации  $a_1$ ,  $a_2$ ,  $a_3$ . Различают мультипликативную и аддитивную модели Хольта-Уинтерса.

где  $a_{0t}$  и  $a_{1t}$  — адаптируемые параметры линейного тренда на  $t$ -м шаге адаптации;  $a_1$ ,  $a_2$ ,  $a_3$  — параметры адаптации;  $f_t$  — адаптируемый параметр сезонных коэффициентов на  $t$ -м шаге адаптации;  $l$  период сезонности.

"Обучение" аддитивной модели Хольта-Уинтерса происходит по формулам:

$$\begin{cases} a_{0t} = a_1 * (x_t - f_{t-l} + (1 - a_t) * (a_{0t-1} + a_{1t-1})) \\ a_{1t} = a_2 * (a_{0t} - a_{0t-1}) + (1 - a_2) * a_{1t-1} \\ f_t = a_3 * (x_t - a_{0t}) + (1 - a_3) * f_{t-l} \end{cases}$$

Прогнозирование в аддитивной модели на  $g$  шагов вперед осуществляется по формуле:

$$x_{t+\tau} = a_{0t} + a_{1t} * \tau + f_{t-l+r}$$

Наиболее сложным моментом при построении модели является определение начальных условий  $a_{00}$ ;  $a_{10}$ ;  $f_{i-l}$  и параметров адаптации модели  $a_1$ ,  $a_2$ ,  $a_3$ . На практике параметры и рассчитываются как коэффициенты линейной регрессии вида  $x_t = a_{00} + a_{10}t$  Начальные значения адаптируемых коэффициентов сезонности  $f_{i-l}$  ( $i = 1, 2, \dots, l$ ) определяются как приростов  $f_t = x_t - \hat{y}_t$ , рассчитанные для каждой одноименной фазы периода (где  $\hat{y}_t$  - расчетные значения линейного тренда определенного для всей обучающей последовательности).

#### 64. Модель Тейла-Вейджа (мультипликативная модель).

Пусть задан временной ряд:  $y_1, \dots, y_t, y_i \in R$

Необходимо решить задачу прогнозирования временного ряда.

Модель Тейла-Вейджа (Theil, Wage) — усложненная модель Хольта, учитывающая сезонность и аддитивный тренд, в отличие от модели Хольта-Уинтерса аддитивно включает линейный тренд, что оправдано при решении некоторых задач.

$$\widehat{y_{t+d}} = a_t + db_t\Theta_t + (d \bmod s) - s$$

$$a_t = a_1(y_t - \Theta_{t-s}) + (1 - a_1)(a_{t-1} + b_{t-1})$$

$$b_t = a_3(a_t - a_{t-1}) + (1 - a_1)b_{t-1}$$

$$\Theta_t = a_2(y_t - a_t) + (1 - a_2)\Theta_{t-s}$$

где  $s$  — период сезонности,  $\Theta$  — сезонный профиль,  $b_t$  — параметр тренда,  $a_t$  — параметр прогноза, очищенный от влияния тренда и сезонности.

Выбирать параметры  $\alpha_1, \alpha_2, \alpha_3 \in (0,1)$  предлагается экспериментально, используя метод минимизации среднеквадратичной ошибки.

### 65. Метод Четверикова.

Исходный ряд  $y_t$  выравнивается по формуле средней хронологической с периодом 1 год, т. е.  $T = 12$ . Не выровненные значения в начале и в конце ряда отбрасывают. Получают предварительную оценку тренда

Этап 1

$$\widetilde{y_t} \equiv f'_t$$

и вычисляют отклонение исходного ряда от выровненного

$$l_{ij} = Y_{ij} - f'_{tj}$$

Этап 2

Для каждого года  $i$  вычисляется среднеквадратическое отклонение

$$\sigma = \sqrt{\frac{\sum_{j=1}^T l_{ij}^2 - (\sum_{j=1}^T l_{ij})^2 / T}{T - 1}}$$

И полученные отклонения нормируются

$$L = l_{ij} / \sigma$$

Этап 3 По нормированным отклонениям вычисляется предварительная средняя сезонная волна

$$S_j^1 = \frac{\sum_{i=1}^m L_{ij}}{m}$$

Этап 4. Средняя предварительная сезонная волна умножается на среднеквадратическое

отклонение каждого года и вычитается из исходного ряда, получается первая оценка тренда

$$f_{ij}^1 = Y_{ij} - S_j^1 \sigma_i$$

Этап 5 Получаемый тренд сглаживают скользящей средней по пяти точкам и получают новую оценку тренда  $f_{ij}^1$ . Чтобы не потерять точки в начале и в конце ряда их сглаживают по трем точкам, причем для крайних точек используют специальные формулы сглаживания

$$f_1^2 = \frac{5f_1^1 + 2f_2^1 - f_3^1}{6}; f_n^2 = \frac{5f_n^1 + 2f_{n-1}^1 - f_{n-2}^1}{6}$$

Этап 6 вычисляют новые отклонения исходного ряда  $y_t$  от тренда  $f_t^2$

Для полученных отклонений вновь выполняются пункты 2 и 3 и получают окончательную среднюю сезонную волну  $S_j^2$

Этап 7 Вычисляют остаточную компоненту  $\epsilon_{ij} = l_{ij}^2 - S_j^2 \sigma_i$  и определяют коэффициент напряженности сезонной волны

$$k_i = \frac{\sum_{j=1}^T l_{ij}^2 \epsilon_{ij}}{\sum_{j=1}^T \epsilon_{ij}^2}$$

## 66. Мультипликативная (аддитивная) модель ряда динамики при наличии тенденции: этапы построения.

Мультипликативная (или аддитивная) модель ряда динамики с учетом тенденции является распространенным подходом в эконометрике для анализа и прогнозирования временных рядов, включая изменение переменной с учетом долгосрочной тенденции. Давайте рассмотрим этапы построения мультипликативной модели ряда динамики при наличии тенденции.

Исходные данные:

временной ряд данных, содержащий значения переменной в последовательных моментах времени. Обычно данные представляют собой наблюдения переменной на равномерных временных интервалах (например, ежемесячно или ежеквартально).

Визуализация:

график временного ряда, чтобы визуально оценить его общую динамику и наличие тренда. На графике можно увидеть, возможно ли наличие долгосрочной тенденции, а также другие особенности временного ряда, такие как сезонность или цикличность.

Декомпозиция временного ряда:

В декомпозицию временного ряда на составляющие компоненты: тренд, сезонность, цикличность и остаток

Для мультипликативной модели каждая компонента будет умножаться между собой:

$$Y(t) = T(t) * S(t) * C(t) * \varepsilon(t) \text{ где}$$

$Y(t)$  - значение переменной в момент времени  $t$ ,

$T(t)$  - трендовая компонента,

$S(t)$  - сезонная компонента,

$C(t)$  - циклическая компонента,

$\varepsilon(t)$  - случайная ошибка.

Оценка параметров

Оцените параметры каждой компоненты модели с использованием соответствующих методов, таких как метод наименьших квадратов или метод максимального правдоподобия.

Для оценки тренда можно использовать линейную или нелинейную регрессию, а для оценки сезонности сезонные индексы.

Циклическую компоненту можно оценить с использованием методов фильтрации, таких как скользящие средние

Прогнозирование:

После оценки параметров модели мультипликативной динамики с учетом тенденции можно использовать ее для прогнозирования будущих значений переменной. Для прогнозирования требуется учитывать прогнозы тренда, сезонности и цикличности, а также случайную ошибку

## 67. Моделирование периодических колебаний (гармоники Фурье).

Моделирование циклических колебаний в целом осуществляется аналогично моделированию сезонных колебаний, поэтому мы рассмотрим только моделирование последних.

Существует несколько подходов к моделированию сезонных (циклических) колебаний:

- расчет значений сезонной компоненты и построение аддитивной или мультипликативной модели временного ряда;
- применение сезонных фиктивных переменных;
- использование рядов Фурье и др.

Поскольку для целей прогнозирования наиболее важным является применение рядов Фурье, остановимся подробнее на этом методе.

С помощью ряда Фурье можно представить периодические (сезонные и циклические) колебания  $f(t)$ , свойственные динамике многих экономических явлений, в виде функции времени  $t$ :

$$f(t) = a_0 + \sum_{k=1}^m (a_k * \cos kt + b_k * \sin kt)$$

параметры ряда Фурье  $a_0, a_k, b_k, k=1, \dots, n$   
 $\cos$  и  $\sin$ - тригонометрические функции (или гармоники ряда Фурье);  
 $k$  - номер гармоники;  $m$  — число гармоник ряда Фурье.

На основе МНК параметры ряда Фурье определяются по следующим формулам:

$$a_0 = \bar{y} = \frac{1}{n} \sum_{t=1}^n y_t \quad a_k = \frac{2}{n} \sum_{t=1}^n y_t * \cos kt \quad b_k = \frac{2}{n} \sum_{t=1}^n y_t * \sin kt$$

## 68. Прогнозирование одномерного временного ряда случайной компоненты (распределение Пуассона).

Распределение Пуассона широко используется для анализа данных, где наблюдаются дискретные события или случайные счетчики.

основные этапы прогнозирования одномерного временного ряда случайной компоненты с распределением Пуассона.

Исходные данные

временной ряд данных, содержащий значения случайной компоненты в последовательных моментах времени.

Распределение Пуассона предполагает, что переменная является дискретной и принимает только неотрицательные целочисленные значения.

Оценка параметров:

Для прогнозирования распределения Пуассона необходимо оценить параметр  $\lambda$  (интенсивность или среднее значение).

Оценка параметра  $\lambda$  может быть выполнена с использованием метода максимального правдоподобия или метода наименьших квадратов.

Формулирование модели:

После оценки параметра  $\lambda$  можно сформулировать модель прогнозирования временного ряда случайной компоненты с распределением Пуассона. –

Модель может быть представлена следующим образом:  $Y(t+1) \sim \text{Poisson}(\lambda)$  где  $Y(t+1)$  - значение случайной компоненты в следующем моменте времени  $t+1$ . 4.

Прогнозирование

Для прогнозирования будущих значений случайной компоненты с распределением Пуассона можно использовать оцененный параметр  $\lambda$ . -

Прогнозируемое значение  $Y(t+1)$  будет являться случайной величиной, распределенной по распределению Пуассона с параметром  $\lambda$ .  $Y(t+1) \sim \text{Poisson}(\lambda)$

Прогнозирование одномерного временного ряда случайной компоненты с распределением Пуассона позволяет оценить будущие значения счетных данных, учитывая их дискретную природу. Этот подход особенно полезен при анализе и прогнозировании экономических и финансовых данных, где наблюдаются счетные события или случайные счетчики

## 69. Функциональные преобразования переменных в линейной регрессионной модели. Метод Зарембки. Особенности применения.

Если стоит выбор между моделью регрессии, внутренне нелинейной и линейной моделью регрессии (или сводящейся к линейному виду), то предпочтение отдаётся линейным формам моделей. Однако многие модели регрессии различной функциональной формы нельзя сравнивать с помощью стандартных критериев (например, сравнение по множественному коэффициенту детерминации, или суммам квадратов отклонений), которые позволили бы подобрать наиболее подходящую модель регрессии.

Функциональное преобразование переменных в линейной регрессионной модели подразумевает выбор наилучшей зависимости

Сравниваются две модели.

$$y_i = a_1 + a_2 X_i + u_i$$

$$\ln(y_i) = \beta_1 + \beta_2 X_i + \ln(u_i)$$

Шаг 1. Вычисляется

$$\overline{Y_{geom}} = \sqrt[n]{y_1 y_2 \dots y_n}$$

Шаг 2 Наблюдения  $y_i$  пересчитываются новые  $y_i^* = \frac{y_i}{\overline{Y_{geom}}}$

Шаг 3. Рассматриваются линейная регрессия с наблюдениями  $y_i^*$  вместо  $y_i$ , и логарифмическая регрессия с наблюдениями  $\ln y_i^*$  вместо  $\ln y_i$ . В остальной модели не меняются (то есть правая часть остается без изменений):

$$y_i = a_1 + a_2 X_i + u_i$$

$$\ln(y_i) = \beta_1 + \beta_2 X_i + \ln(u_i)$$

Находим остаточные суммы квадратов остатков для полученных вспомогательных регрессий ESS1 и ESS2;

Шаг 4. Составляем статистика  $Z = \left| \frac{n}{2} \ln \left( \frac{ESS1}{ESS2} \right) \right|$  и сравнивается с

табличным распределением Хи-квадрат с одной степенью свободы.

Если  $Z > \chi^2$  то гипотеза относительно того, что модели (1) и (2) не имеют статистически значимых различий, отвергается, то есть выбираем модель полулогарифмическую, если  $Z < \chi^2$  — выбираем линейную модель.

Подходит для выбора между линейной и полулогарифмической моделью

## 70. Функциональные преобразования переменных в линейной регрессионной модели. Тест Бокса-Кокса. Особенности применения.

Функциональная зависимость - форма зависимости  $Y$  от каждой объясняющей переменной. Функциональное преобразование переменных в линейной регрессионной модели подразумевает выбор наилучшей зависимости. Тест Бокса-Кокса основывается на утверждении о том, что  $(y-1)$  и  $\ln y$  являются частными случаями функции вида

$$F = \frac{y^\lambda - 1}{\lambda}$$

Если параметр  $\lambda = 1$  то данная формула принимает вид  $F = y - 1$

Если параметр  $\lambda \rightarrow 0$  то функция принимает вид  $F = \ln(y)$

Шаг 1. Преобразуем зависимую переменную по методу П. Зарембеки:

$$y_i^* = \frac{y_i}{\sqrt[n]{y_1 y_2 \dots y_n}}$$

Шаг 2. Рассчитываются новые переменные (преобразование Бокса-Кокса):

$$y_{i(B-C)} = \frac{y_i^{8\lambda}}{\lambda} \text{ и } X_{i(B-C)} = \frac{x_i^\lambda}{\lambda}$$

Шаг 3. Оценивается линейная модель регрессии с использованием масштабированных значений при  $\lambda$  от 0 до 1

$$y_{i(B-C)} = a_1 + a_2 X_{i(B-C)} + u_i$$

Шаг 4. Выбирается минимальное значение суммы квадратов остатков ESS, выбирается одна из крайних регрессий, к которой ближе точка минимума.

### Особенности применения

Подходит для выбора между линейной и логарифмической моделью

## 71. Функциональные преобразования переменных в линейной регрессионной модели. Критерий Акаике и Шварца. Особенности применения.

Статистика Акаике:

$$AIC = \ln\left(\frac{ESS_k}{n}\right) + \frac{2k}{n} + 1 + \ln(2\pi)$$

При увеличении объясняющих переменных первое слагаемое в правой части уменьшается, второе – увеличивается. Среди нескольких альтернативных моделей (полной и редуцированной) предпочтение отдается модели, с наименьшим значением статистики AIC, в которой достигается определенный компромисс между величиной остаточной суммы квадратов и количеством объясняющих переменных. Особенности критерия: Штрафование числа параметров ограничивает значительный рост сложности модели, проверка критерия является трудоемкой операцией, может сравнивать модели только с выборками равного размера, порядок выбора моделей неважен.

Статистика Шварца:

$$SC = \ln\left(\frac{ESS_k}{n}\right) + \frac{k \ln(n)}{n} + 1 + \ln(2\pi)$$

При увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а второе – увеличивается. Среди нескольких альтернативных моделей (полной и редуцированной) предпочтение отдается модели с наименьшим значением статистики Шварца.

Используется при длинных выборках данных. Отличается следующими признаками: Из двух моделей предпочтительно выбрать с меньшим значением байесовского критерия.

Байесовский критерий представляет собой возрастающую функцию от числа параметров модели и от остаточной суммы квадратов ошибок модели.

Изменение зависимых переменных и увеличение числа наблюдаемых увеличивает байесовский критерий, в то же время уменьшение критерия означает уменьшение размерности модели.

Используется при длинных выборках данных.

## 72. Функциональные преобразования переменных в линейной регрессионной модели. Тест Бера. Особенности применения.

Тест для выбора между полулогарифмической и линейной моделью

$$H_0: \ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

$$H_1: Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

Шаг 1: найдем оцененные значения зависимой переменной в каждой модели  $\hat{y}$ ,  $\widehat{\ln y}$

Шаг 2: составим вспомогательные регрессии

$$e^{\widehat{\ln Y}} = \beta_0 + \sum_{i=1}^k \beta_i x_i + v_1$$

$$\ln \hat{Y} = \beta_0 + \sum_{i=1}^k \beta_i x_i + v_2$$

Шаг 3: составим вспомогательные регрессии

$$\ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_1 \hat{v}_1 + \varepsilon_1$$

$$Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_2 \hat{v}_2 + \varepsilon_2$$

Шаг 4: используется t тест. Если  $\theta_1 = 0$ , не отвергается, а  $\theta_2 = 0$  отвергается, выбирается полулогарифмическая модель. Если  $\theta_2 = 0$ , не отвергается, а  $\theta_1 = 0$  отвергается, выбирается линейная модель

### 73. Функциональные преобразования переменных в линейной регрессионной модели. Тест МакАлера. Особенности применения.

Тест для выбора между полулогарифмической и линейной моделью

$$H_0: \ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$
$$H_1: Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

Шаг 1: найдем оцененные значения зависимой переменной в каждой модели  $\hat{y}$ ,  $\widehat{\ln y}$

Шаг 2: составим вспомогательные регрессии

$$e^{\widehat{\ln Y}} = \beta_0 + \sum_{i=1}^k \beta_i x_i + v_1$$
$$\ln \hat{Y} = \beta_0 + \sum_{i=1}^k \beta_i x_i + v_2$$

Шаг 3: составим вспомогательные регрессии

$$\ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_1 \hat{v}_1 + \varepsilon_1$$
$$Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_2 \hat{v}_2 + \varepsilon_2$$

Шаг 4: используется t тест. Если  $\theta_1 = 0$ , не отвергается, а  $\theta_2 = 0$  отвергается, выбирается полулогарифмическая модель. Если  $\theta_2 = 0$ , не отвергается, а  $\theta_1 = 0$  отвергается, выбирается линейная модель

#### 74. Функциональные преобразования переменных в линейной регрессионной модели. Тест МакКиннона. Особенности применения.

Тест для выбора между полулогарифмической и линейной моделью

$$H_0: \ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$
$$H_1: Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

Шаг 1: найдем оцененные значения зависимой переменной в каждой модели  $\hat{y}$ ,  $\widehat{\ln y}$

Шаг 2: составим вспомогательные регрессии

$$\ln y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_1 [\hat{Y} - \exp(\widehat{\ln Y})] + \varepsilon_1$$
$$Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_2 [\widehat{\ln Y} - \ln \hat{Y}] + \varepsilon_2$$

Используем t тест. Если  $\theta_1 = 0$ , не отвергается, а  $\theta_2 = 0$  отвергается, выбирается полулогарифмическая модель. Если  $\theta_2 = 0$ , не отвергается, а  $\theta_1 = 0$  отвергается, выбирается линейная модель

## 75. Функциональные преобразования переменных в линейной регрессионной модели. Тест Уайта. Особенности применения.

Тест для выбора между полулогарифмической и линейной моделью

$$H_0: \ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

$$H_1: Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

Шаг 1: найдем оцененные значения зависимой переменной в каждой модели  $\hat{y}$ ,  $\widehat{\ln y}$

Шаг 2: составим вспомогательные регрессии

$$\ln y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_1 [\hat{Y} - \exp(\widehat{\ln Y})] + \varepsilon_1$$

$$Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_2 [\widehat{\ln Y} - \ln \hat{Y}] + \varepsilon_2$$

Используем t тест. Если  $\theta_1 = 0$ , не отвергается, а  $\theta_2 = 0$  отвергается, выбирается полулогарифмическая модель. Если  $\theta_2 = 0$ , не отвергается, а  $\theta_1 = 0$  отвергается, выбирается линейная модель

76. Функциональные преобразования переменных в линейной регрессионной модели. Тест Дэвидсона. Особенности применения.

Тест для выбора между полулогарифмической и линейной моделью

$$H_0: \ln Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$
$$H_1: Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

Шаг 1: найдем оцененные значения зависимой переменной в каждой модели  $\hat{y}$ ,  $\widehat{\ln y}$

Шаг 2: составим вспомогательные регрессии

$$\ln y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_1 [\hat{Y} - \exp(\widehat{\ln Y})] + \varepsilon_1$$
$$Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \theta_2 [\widehat{\ln Y} - \ln \hat{Y}] + \varepsilon_2$$

Используем t тест. Если  $\theta_1 = 0$ , не отвергается, а  $\theta_2 = 0$  отвергается, выбирается полулогарифмическая модель. Если  $\theta_2 = 0$ , не отвергается, а  $\theta_1 = 0$  отвергается, выбирается линейная модель

## 77. Модели с распределенными лаговыми переменными.

Модель регрессии по временным предам с лаговыми объясняющими переменными – модель с распределёнными лагами

$$y_t = a + b_0x_t + b_1x_{t-1} + \dots + b_kx_{t-k} + \xi_t$$

Подразделяются на два типа:

с конечными лагами  $y_t = a + b_0x_t + b_1x_{t-1} + \dots + b_kx_{t-k} + \xi_t$

с бесконечным числом лагов  $y_t = a + b_0x_t + b_1x_{t-1} + b_2x_{t-2} + \dots + \xi_t$

Любая сумма коэффициентов  $\sum_{j=0}^h b_j$ , с  $h < k$  – промежуточный

мультипликатор, а сумма всех коэффициентов  $\sum_{j=0}^k b_j$  – долгосрочный

мультипликатор, который характеризует общее изменение  $y$  через  $k$  интервалов времени под воздействием изменения  $x$  в момент  $t$  на 1 ед.

Относительные коэффициенты модели  $\beta_j$  можно использовать как весовые коэффициенты для расчета средней величины лага по средней арифметической, где  $j$  - величина лага

$$\bar{j} = \sum_{j=0}^k j\beta_j$$

Величина  $\bar{j}$  показывает средний интервал времени, в течение которого будет происходить изменение зависимой переменной  $y$  под воздействием изменения объясняющей переменной  $x$  в момент времени  $t$ . Чем меньше величина среднего лага, тем быстрее реагирует результат  $y$  на изменения  $x$ . Кроме среднего лага можно рассчитать медианный лаг  $j_M$ , т. е. тот период времени, в течение которого с момента времени  $t$  будет реализована половина общего эффекта воздействия объясняющей переменной  $x$  на результат  $y$ .

$$\sum_{j=0}^{M_e} \beta_j = 0,5$$

## 78. Оценка моделей с лагами в независимых переменных.

### Преобразование Койка.

Модели с распределёнными лагами (то есть лаговыми объясняющими переменными) подразделяются модели с конечным и бесконечным числом лагов. Для модели с конечным числом лагов при правильной ее спецификации может быть оценена обычным МНК. Но в таком случае возможна мультиколлинеарность факторов и автокорреляция остатков. Поэтому нередко для оценки параметров модели с распределённым конечным числом лагов используются специальные методы преобразования, как и для модели с бесконечным числом лагов. Разработаны разные методы оценивания параметров модели с распределёнными лагами, которые учитывают характер распределения коэффициентов регрессии при лаговых объясняющих переменных. Иными словами, методы оценивания параметров модели с распределёнными лагами основаны на изучении структуры лага. Так, предполагая полиномиальное распределение лаговых коэффициентов, используют метод Алмон, а при гипотезе геометрической прогрессии для лаговых коэффициентов применяется преобразование Койка. Это преобразование используется для моделей с бесконечным числом лагов.

$$y_t = a + b_0x_t + b_1x_{t-1} + b_2x_{t-2} + \dots + \xi_t(1)$$

Предполагая, что в модели все лаговые коэффициенты имеют одинаковый знак и уменьшаются в геометрической прогрессии Койк предложил для оценки параметров модели следующий процедуру.

Пусть имеется модель  $y_t = a + b_0x_t + \lambda b_0x_{t-1} + \lambda^2 b_0x_{t-2} + \dots + U_t$ , где предположено, что убывание наглых коэффициентов геометрической прогрессии происходит сразу же, а не через интервал времени.

Построить модель для момента времени  $t-1$

$$y_{t-1} = a + b_0x_{t-1} + \lambda b_0x_{t-2} + \lambda^2 b_0x_{t-3} + \dots + \xi_{t-1}(3)$$

умножить уравнение на  $\lambda$

$$\lambda y_{t-1} = \lambda a + b_0\lambda x_{t-1} + \lambda^2 b_0x_{t-2} + \lambda^3 b_0x_{t-3} + \dots + \lambda \xi_{t-1}(4)$$

Вычесть из уравнения 1 уравнение 4

$$y_t - \lambda y_{t-1} = (1 - \lambda)a + b_0x_t + (\xi_t - \lambda \xi_{t-1})$$

после преобразования получить

$$y_t = (1 - \lambda)a + b_0x_t + \lambda y_{t-1} + U_t$$

Где  $U_t = \xi_t - \lambda \xi_{t-1}$

Преобразование может быть использовано и при решении модели, когда несколько первых коэффициентов остаются свободными, а для оставшихся лагов реализуются данное преобразование. Преобразование приводит к существенным упрощениям, ибо вместе с уменьшением числа оцениваемых параметров устраняется и проблемы мультиколлинеарности факторов. Модель позволяет анализировать краткосрочный ( $b_0$ ) и долгосрочный (сумма коэффициентов регрессии, представляющая собой сумму геометрической прогрессии) мультипликаторы. Оценивание параметров реализуются методом максимального правдоподобия или инструментальными переменными. Остатки могут быть автокоррелированы.

## 79 Полиномиально распределенные лаги Алмон 79

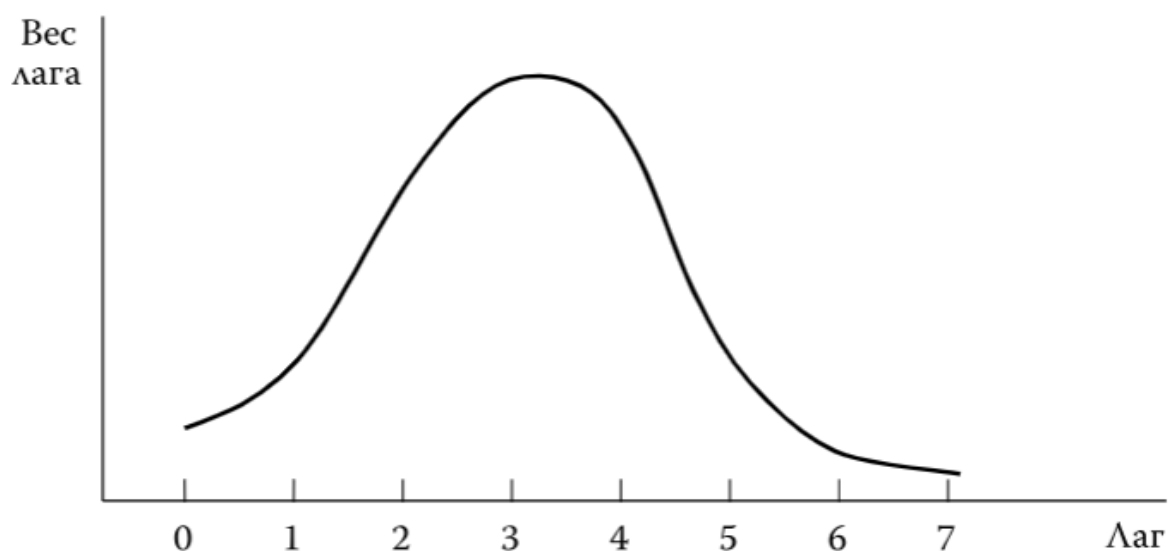
В 1965 г. Ш. Алмон предложила способ оценки параметров модели с распределенными лагами на основе гипотезы о том, что лаговые коэффициенты регрессии аппроксимируются полиномом соответствующей степени от величины лага. Это значит, что в модели  $y_t = \alpha + b_0x_t + b_1x_{t-1} + \dots + b_kx_{t-k} + \varepsilon_t$  параметр  $b_j$  рассматривается как функция:  $b_j = c_0 + c_1j + c_2j^2 + \dots + c_mj^m$ . При этом априори выдвигается предположение о степени полинома. Как правило, используется многочлен невысокой степени ( $m \leq 4$ ).

В общем виде при степени полинома  $m$  модель регрессии с распределенными лагами примет вид:

$$y_t = \alpha + c_0(x_t + x_{t-1} + \dots + x_{t-k}) + c_1(x_{t-1} + 2x_{t-2} + 3x_{t-3} \dots + kx_{t-k}) + \dots + c_m(x_{t-1} + 2^mx_{t-2} + 3^mx_{t-3} \dots + k^mx_{t-k}) + \varepsilon_t$$

Или  $y_t = \alpha + c_0z_0 + c_1z_1 + c_2z_2 + \dots + c_mz_m + \varepsilon_t$

Как видим, в данной модели переменные  $z_1, z_2, \dots, z_m$  представляют собой линейную комбинацию переменных  $x_t$  и  $k$  лаговых переменных, веса при которых подчиняются полиномиальному распределению.



## 80 Авторегрессионные модели 80

Модель временного ряда, в которой его текущее значение линейно зависит от предыдущих (ретроспективных) значений этого же ряда.

Линейная зависимость означает, что текущее значение равно взвешенной сумме нескольких предыдущих значения ряда, т.е.

$$Y(t) = C + b_1 Y(t-1) + b_2 Y(t-2) + \dots + b_n Y(t-n) + \varepsilon(t) = C + \sum_{i=1}^n b_i Y_{t-n} + \varepsilon(t),$$

где  $C$  — константа, которую для простоты часто полагают равной 0;  $n$  — число ретроспективных значений ряда, учитываемых в модели (порядок модели);  $b_i$  — коэффициенты (параметры) модели, которые требуется оценить при ее построении;  $\varepsilon(t)$  — случайная составляющая, отражающая вероятностный характер модели.

Если временной ряд представляет собой ежедневные продажи, то  $Y(t)$  — продажи сегодня,  $Y(t-1)$  — продажи, которые были вчера,  $Y(t-2)$  — позавчера и т.д.,  $\varepsilon(t)$  — учитывает влияние на продажи случайных факторов, которые невозможно учесть в модели (например, погоду или колебания курса доллара). Возможно использование и других шкал наблюдений — еженедельные, ежеквартальные и т.д.

Таким образом, зная параметры модели и соответствующие ретроспективные значения временного ряда, мы можем предсказать его будущие значения. Поэтому основное назначение авторегрессионной модели — прогнозирование. Кроме этого, с ее помощью можно производить анализ временных рядов — выявлять тенденции, сезонность и другие особенности.

## 81 Авторегрессионные модели с распределёнными лагами 81

Модель авторегрессии и распределённого лага (ADL-модель, [англ. autoregressive distributed lags](#)) — модель [временного ряда](#), в которой текущие значения ряда зависят как от прошлых значений этого ряда, так и от текущих и прошлых значений других временных рядов.

Модель ADL(p,q) с одной экзогенной переменной имеет вид:

$$y_t = a_0 + \sum_{i=1}^p a_i y_{t-i} + \sum_{j=0}^q b_j x_{t-j} + \varepsilon_t$$

Модель ADL(p,0) — это [модель авторегрессии AR\(p\)](#) (в общем случае, возможно с [экзогенной](#) переменной без лагов), а модель ADL(0,q) — это модель распределённого лага DL(q).

Модель обобщается на случай нескольких экзогенных переменных  $x$ . В этом случае возможно обозначение модели ADL(p,q1,q2,...,qk), где  $k$  — количество экзогенных переменных,  $q_i$  — количество лагов  $i$ -ой переменной, входящих в модель. В общем случае, можно считать, что все экзогенные переменные включены в модель с одинаковым количеством лагов, а исключение какого-либо лага некоторых переменных означает лишь ограничение на модель. Поэтому иногда используют обозначение ADL(p,q,k),  $k$  — количество экзогенных переменных,  $q$  — количество лагов. Наложение ограничений на коэффициенты этой модели приводит к тем или иным вариациям. В таком обозначении, классическая модель ADL(p,q) будет обозначаться как ADL(p,q,1).

На практике для оценки подобных моделей часто используют [методологию Бокса-Дженкинса](#) для оценки авторегрессии и специальные приёмы для упрощения оценки [распределённого лага](#)

Рассмотрим модель :

$$y_t = a_0 + a_1 y_{t-1} + b_0 x_t + b_1 x_{t-1} + e_t$$

Таким образом краткосрочная зависимость выражается коэффициентом  $b_0$  реакции на изменение фактора по сравнению с прошлым периодом.

## 82 Стационарные временные ряды. Определения стационарности, лаговой переменной, автоковариационной функции временного ряда, автокорреляционной функции, коррелограммы, коэффициенты корреляции между разными элементами стационарного временного ряда с временным лагом

Определение. Временной ряд  $\{x_t\}$  называется стационарным (в широком смысле), если

$$E_{x_t} \equiv \text{const} (\text{среднее постоянно во времени});$$

$$\text{cov}(x_t, x_{t+h}) = \gamma(h) (\text{ковариация зависит только от лага } h)$$

Понятие стационарного временного ряда означает, что его среднее значение не изменяется во времени, т.е. временной ряд не имеет тренда.

Кроме того, ковариация между разными элементами временного ряда (как между случайными величинами) зависит только от того, насколько сильно они отдалены друг от друга во времени. Величина, характеризующая разницу во времени между элементами временного ряда, называется лаговой переменной или запаздыванием. Так как

$$\gamma(0) = \text{cov}(x_t, x_t) = \text{Var}(x_t)$$

то дисперсия стационарного временного ряда также не меняется со временем.

**Определение.** Функция  $\gamma(h)$  как функция от лаговой переменной, называется автоковариационной функцией временного ряда.

Она определена как для положительных, так и для отрицательных лагов  $h$ . Так как

$$\gamma(-h) = \text{cov}(x_t, x_{t-h}) = \text{cov}(x_{t-h}, x_t) = \text{cov}(x_t, x_{t+h}) = \gamma(h)$$

то  $\gamma(h)$  — четная функция. Для произвольных моментов времени  $t$  и  $s$  очевидно равенство

$$\text{cov}(x_t, x_s) = \gamma(t - s)$$

Вычислим теперь коэффициенты корреляции между разными элементами стационарного временного ряда с временным лагом  $h$ :

$$\text{corr}(x_t, x_{t+h}) = \frac{\text{cov}(x_t, x_{t+h})}{\sqrt{\text{Var}(x_t) * \text{Var}(x_{t+h})}} = \frac{\gamma(h)}{\sqrt{\gamma(0) * \gamma(0)}} = \frac{\gamma(h)}{\gamma(0)}$$

Также, как и в случае коэффициента ковариации, коэффициент корреляции между разными элементами стационарного временного ряда зависит только от лага между ними. Например,  $\text{corr}(x_1, x_3) = \text{corr}(x_7, x_9) = \text{corr}(x_{15}, x_{17})$  и  $\text{corr}(x_2, x_7) = \text{corr}(x_{10}, x_{15}) = \text{corr}(x_{24}, x_{29})$ .

Определение. Функция  $p(h) = \text{corr}(x_t, x_{t+h})$  называется автокорреляционной функцией (autocorrelation function, ACF) стационарного временного ряда.

Очевидно, что она также является четной функцией лаговой переменной и  $p(0) = 1$ . Для коэффициента автокорреляции очевидно:

$$\text{corr}(x_s, x_t) = p(s - t)$$

Определение. Коррелограммой стационарного временного ряда называется график функции  $p(h)$ .

Наряду с автокорреляционной функцией ещё рассматривается частная автокорреляционная функция, представляющая собой частный коэффициент корреляции между уровнями временного ряда  $x_t$  и  $x_{t+h}$  при исключении влияния промежуточных уровней  $x_{t+1}, \dots, x_{t+h-1}$ .

### 83 Стационарные временные ряды. Определения частной автокорреляционной функции, белого шума, автоковариационная функция для белого шума, ACF для белого шума, частная автокорреляционная функция для белого шума 83

Определение. Частная автокорреляционная функция

$$p_{part}(h) = corr(x_t, x_{t+h} | x_{t+1}, \dots, x_{t+h-1})$$

Очевидно  $p_{part}(0) = 1, p_{part}(1) = \rho(1)$

При общих условиях

$$\lim_{h \rightarrow \infty} p_{part}(h) = 0$$

Рассмотрим основной пример стационарного временного ряда.

Определение. Ряд  $u_t$ , называется белым шумом (white noise), если

$$1. E u_t = 0,$$

$$2. Var(u_t) = E u_t^2 \equiv \sigma^2$$

$$3. cov(u_t, u_{t+h}) = E(u_t, u_{t+h}) = 0 \text{ при } h \neq 0$$

Обозначение  $u_t \sim WN(0, \sigma^2)$

Белый шум удобно рассматривать как экзогенный шок (не коррелирующий с прошлым). Он используется для построения моделей стационарных рядов.

Определение. Если  $u_t \sim N$ , то говорят о гауссовском белом шуме

Автокорреляционная функция для белого шума:

$$\gamma(h) = \begin{cases} \sigma^2, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

ACF для белого шума:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

Это означает, что ряд “мгновенно забывает” прошлые значения.

Частная автокорреляционная функция для белого шума:

$$p_{part}(h) = \begin{cases} 1, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

## 84 Модели стационарных временных рядов: модель (классический вид и через лаговый оператор).

### Авторегрессионный многочлен, авторегрессионная часть и часть скользящего среднего 84

Общая смешанная модель ARMA авторегрессии-скользящего среднего

Для удобства представления различных моделей часто используется (формальный) лаговый оператор  $L$ :

$$L(x_t) = x_{t-1}$$

Далее

$$L^2(x_t) = L(L(x_t)) = L(x_{t-1}) = x_{t-2}$$

Следовательно,

$$L^k(x_t) = x_{t-k}$$

и формально положим

$$L^0(x_t) = x_t$$

Начнём сразу с общего вида модели ARMA(p,q)

$$x_t = \mu + \sum_{j=1}^p \phi_j x_{t-j} + u_t + \sum_{s=1}^q \theta_s u_{t-s}$$

$$u_t \sim WN(0, \sigma_u^2), \phi_p, \theta_q \neq 0$$

Замечание. Проинтерпретировать модель можно следующим образом: текущее значение зависит от прошлых значений до лага  $p$  и от текущего и прошлых внешних шоков до лага  $q$ . Коэффициенты такой модели в общем случае не имеют экономической интерпретации.

Запишем (1.1) используя лаговый оператор  $L$ :

$$x_t = \mu + \sum_{j=1}^p \phi_j L^j x_t + u_t + \sum_{s=1}^q \theta_s L^s u_t$$

Перепишем в виде:

$$\left(1 - \sum_{j=1}^p \phi_j L^j\right) x_t = \mu + \left(1 + \sum_{s=1}^q \theta_s L^s\right) u_t$$

Теперь введём два многочлена степени  $p$  и  $q$ :

$$\phi(z) = 1 - \sum_{j=1}^p \phi_j z^j = 1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p$$

$$\theta(z) = 1 + \sum_{s=1}^q \theta_s z^s = 1 + \theta_1 z + \theta_2 z^2 + \dots + \theta_q z^q$$

Тогда модель (1.1) формально можно записать

$$\phi(L)x_t = \mu + \theta(L)u_t$$

Определение. Многочлен  $\phi(z)$  называется авторегрессионным многочленом.

Определение.  $\phi(L)x_t$ , называется авторегрессионной частью модели ARMA, а  $\theta(L)u_t$  - частью скользящего среднего.

Утверждение. Модель ARMA определяет стационарный ряд  $\Leftrightarrow$  выполнено условие стационарности: все корни (в том числе из  $\mathbb{C}$ ) авторегрессионного многочлена

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$$

по модулю больше единицы.

85 Модели стационарных временных рядов: модель .

Доказательство утверждения: Модель ARMA(1, q) стационарна тогда и только тогда, когда 85

Пример. Рассмотрим ARMA(1, q):

$$x_t = \mu + \phi x_{t-1} + u_t + \sum_{s=1}^q \theta_s u_{t-s}, \quad \phi, \theta_q \neq 0$$

Тогда  $\phi(z) = 1 - \phi z$  и его корень  $z_0 = 1/\phi$ . Так как  $|z_0| > 1 \Leftrightarrow |\phi| < 1$ , то это и будет условием стационарности для этого ряда. Утверждение. Модель ARMA(1,q) стационарна тогда и только тогда, когда  $|\phi| < 1$ .

## 86 Модели стационарных временных рядов: Модель, Среднее, дисперсия и ACF для MA(q). Модель 86

Рассмотрим частные случаи общей модели ARMA:

MA(q) = ARMA(0,q) – модель скользящего среднего;

Модель MA(q)

Модель MA(q) (q –порядок лага)

$$x_t = \mu + u_t + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}, \quad u_t \sim WN(0, \sigma_u^2) \quad \theta_q \neq 0$$

Временной ряд учитывает только внешние шоки до порядка q. Модель задаёт стационарный ряд  $x_t$  при любых  $\{\theta_j\}_{j=1}^q$ , так как  $\phi(z) = 1$  не имеет корней. Используя лаговый оператор её можно записать так

$$x_t = \mu + \theta(L)u_t$$

Среднее, дисперсия и ACF для MA(q)

Так как  $u_t \sim WN(0, \sigma_u^2)$ , то

$$Ex_t = \mu + Eu_t + \theta_1 Eu_{t-1} + \dots + \theta_q Eu_{t-q} = \mu$$

Далее

$$Var(x_t) = Var(\mu + u_t + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}) = (1 + \theta_1^2 + \dots + \theta_q^2) \sigma_u^2$$

Что касается ACF, то  $p(h) = 0$  при  $|h| > q$ , т.е ряд “забывает” прошлые значения с лагами больше порядка модели.

Если рассмотреть модель MA ( $\infty$ )

$$x_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \dots = \mu + u_t + \sum_{j=1}^{\infty} \theta_j u_{t-j}$$

$$u_t \sim WN(0, \sigma_u^2)$$

В которой ряд “помнит” все прошлые шоки, то для неё

$$Ex_t = \mu, Var(x_t) = (1 + \sum_{j=1}^{\infty} \theta_j^2) \sigma_u^2$$

## 87 модели стационарных временных рядов: модель AR(p)

Доказательство утверждения: Модель AR(p) определяет стационарный ряд  $\Leftrightarrow$  выполнено условие стационарности: все корни многочлена по модулю больше единицы.

Временной ряд  $\{x_t\}$  называется стационарным (в широком смысле), если  $E x_t \equiv \text{const}$  (среднее постоянно во времени);

$\text{cov}(x_t, x_{t+h}) = \gamma(h)$  (ковариация зависит только от лага  $h$ ).

Понятие стационарного временного ряда означает, что его среднее значение не изменяется во времени, т.е. временной ряд не имеет тренда.

Кроме того, ковариация между разными элементами временного ряда (как между случайными величинами) зависит только от того, насколько сильно они отдалены друг от друга во времени.

Модель AR(p)

$$x_t = \mu + \varphi_1 x_{t-1} + \dots + \varphi_p x_{t-p} + u_t \quad u_t \sim WN(0, \sigma_u^2), \varphi_p \neq 0$$

Видно, что текущее значение в этой модели зависит от прошлых значений до лага  $p$  и от текущего внешнего шока.

**Утверждение.** Модель AR(p) определяет стационарный ряд  $\Leftrightarrow$  выполнено условие стационарности: все корни (в том числе из  $\mathbb{C}$ ) многочлена  $\varphi(z)$  по модулю больше единицы.

$$\varphi(z) = 1 - \varphi_1 z - \dots - \varphi_p z^p$$

Возьмем модель AR(1)

$$x_t = \mu + \varphi x_{t-1} + \mu_t, \quad \varphi \neq 0$$

Тогда  $\varphi(z) = 1 - \varphi z$  и его корень  $z_0 = 1/\varphi$ . Так как  $|z_0| > 1 \Leftrightarrow |\varphi| < 1$ , то это будет условием стационарности для этого ряда.

88. Прогнозирование для модели ARMA. Условия прогнозирования. Периоды прогнозирования. Информативность прогнозов.

Начнём сразу с общего вида модели ARMA( $p, q$ )

$$x_t = \mu + \sum_{j=1}^p \phi_j x_{t-j} + u_t + \sum_{s=1}^q \theta_s u_{t-s}$$

$$u_t \sim \text{WN}(0, \sigma_u^2), \quad \phi_p, \theta_q \neq 0.$$

Пусть известны значения ряда  $x_t$  и шоки  $u_t$  до периода  $T$ . Обозначим  $\Omega_T = \{x_t, u_t | t \leq T\}$  и на основе  $\Omega_T$  будем строить «оптимальный» прогноз  $\hat{x}_{T+\tau} = \hat{x}_{T+\tau} | \Omega_T$  на период  $T + \tau$ .

Оптимальность прогноза будем понимать в смысле среднего квадратичного отклонения

$$\min E(\hat{x}_{T+\tau} - x_{T+\tau})^2$$

Для общего случая запишем правило последовательного построения оптимального прогноза на  $\tau$  шагов:

записываем ARMA-формулу для  $x_{T+\tau}$ ;

“отбрасываем” будущие шоки  $u_{T+\tau}, \dots, u_{T+1}$ ;

заменяем будущие значения  $x_{T+\tau-1}, \dots, x_{T+1}$  на их прогнозы, полученные на предыдущих шагах.

Модели стационарных рядов (ARMA) дают только краткосрочные “содержательные” прогнозы. Долгосрочные прогнозы не являются информативными.

## 89. Оценка и тестирование модели: Предварительное тестирование на белый шум.

Предварительное тестирование на белый шум

Для построения моделей необходимо провести предварительный анализ стационарного ряда. Для этого тестируется гипотеза:

$$H_0: x_t \sim WN(\mu, \sigma^2)$$

Точнее тестируется (для некоторого  $K > 0$ ) гипотеза:

$$H'_0: \rho(1) = \dots = \rho(K) = 0$$

Обычно используется две тестовых  $Q$ -статистики, которые автоматически вычисляются в компьютерных пакетах:

Box G.E.D., Pierce D.A., 1970 (устаревшая):

$$Q_{BP} = n \sum_{h=1}^K \hat{p}^2(h)$$

Ljung G.M., Box G.E.D., 1978:

$$Q = n(n+2) \sum_{h=1}^K \frac{\hat{p}^2(h)}{n-h}$$

.

Критическое значение имеет хи-квадрат распределения  $\chi^2_{cr} = \chi^2_{K}(\alpha)$ , поэтому для проверки гипотезы применяем следующее статистическое правило:

При  $Q > \chi^2_{cr}$  отвергаем  $H'_0$  (и  $H_0$ ).

При  $Q < \chi^2_{cr}$  не отвергаем  $H'_0$ .

Или альтернативно делаем вывод с помощью  $P$ -значения.

Замечание. Применение  $Q$ -статистик оправдано только для больших выборок. Это асимптотический тест!

## 90. Оценка модели и тестирование гипотез временного ряда.

Оценивание модели ARMA(p, q) происходит с помощью метода максимального правдоподобия (из-за нелинейности в части МА).

Тестирование гипотез о значимости коэффициентах проводится стандартным для метода максимального правдоподобия путём:

Проверка значимости коэффициента: тестовая статистика  $z = \frac{\hat{\phi}_j}{s_j}$  или  $z = \frac{\hat{\theta}_l}{s_l}$

Проверка совместной значимости:  $LR$  или  $W$  статистики.

Оценки коэффициентов  $\phi_1, \dots, \phi_p$  можно получить как решения системы уравнений Юла–Уолкера

$$\begin{cases} \hat{p}(1) = \phi_1 \hat{p}(0) + \phi_2 \hat{p}(1) + \dots + \phi_p \hat{p}(p-1) \\ \hat{p}(2) = \phi_1 \hat{p}(1) + \phi_2 \hat{p}(0) + \dots + \phi_p \hat{p}(p-2) \\ \dots \\ \hat{p}(p) = \phi_1 \hat{p}(p-1) + \phi_2 \hat{p}(p-2) + \dots + \phi_p \hat{p}(0) \end{cases},$$

а оценка коэффициента  $\mu$  равна

$$m = \bar{x}(1 - \hat{\phi}_1 - \dots - \hat{\phi}_p).$$

## 91. Информационные критерии для сравнения моделей и выбора порядка временного ряда: Акаике, Шварца, Хеннана-Куина.

### Условия их применения.

Информационные критерии Акаике (AIC), Шварца (BIC) и Ханнана-Куина (HQIC) являются статистическими мерами для сравнения различных моделей и выбора порядка временного ряда в анализе временных рядов.

Критерий Акаике (AIC):

$$AIC = -\frac{2\ln L}{n} + \frac{2k}{n}$$

где  $k$  — число параметров в статистической модели,  $L$  — максимизированное значение функции правдоподобия модели.

Условия применения: Критерий Акаике может быть использован для сравнения моделей временных рядов, построенных с использованием одного и того же набора данных.

Интерпретация: Модель с более низким значением AIC считается более предпочтительной. AIC учитывает баланс между точностью модели и ее сложностью. Модели с более высокой точностью и более низкой сложностью получают более низкое значение AIC.

Критерий Шварца (BIC):

$$SC = -\frac{2\ln L}{n} + \frac{k\ln(n)}{n}$$

Условия применения: Критерий Шварца также может быть использован для сравнения моделей временных рядов на основе одного набора данных.

Интерпретация: Модель с более низким значением BIC считается более предпочтительной. BIC принимает во внимание точность модели и ее сложность, но штрафует более сложные модели сильнее, чем AIC. По сравнению с AIC, BIC более склонен выбирать более простые модели.

Критерий Ханнана-Куина (HQIC):

$$HQ = -\frac{2\ln L}{n} + \frac{2k\ln\ln(n)}{n}$$

Условия применения: Критерий Ханнана-Куина также используется для сравнения моделей временных рядов.

Интерпретация: Модель с более низким значением HQIC считается более предпочтительной. HQIC подобен BIC в том смысле, что он штрафует более сложные модели, но обычно дает более высокие значения, чем BIC. Все три критерия предоставляют меру относительного качества моделей, и каждый из них имеет свои преимущества и недостатки. Выбор между ними зависит от конкретного контекста и целей исследования. Некоторые исследователи могут предпочитать AIC, некоторые — BIC или HQIC, в зависимости от своих предположений о модели и приоритетах.

## 92. Проверка адекватности модели: тесты на автокорреляцию временного ряда Дарбина-Уотсона, Льюинга-Бокса.

Тест Дарбина-Уотсона (Durbin-Watson test) и тест Льюинга-Бокса (Ljung-Box test) являются двумя распространенными методами для проверки адекватности моделей временных рядов, особенно моделей авторегрессии (AR) или скользящего среднего (MA).

Тест Дарбина-Уотсона:

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

Условия применения: Тест Дарбина-Уотсона применяется для проверки автокорреляции первого порядка в остатках модели.

Интерпретация: Тест вычисляет статистику Дарбина-Уотсона, которая принимает значения от 0 до 4. Значение близкое к 2 указывает на отсутствие автокорреляции в остатках (адекватность модели), значения меньше 2 указывают на положительную автокорреляцию, а значения больше 2 указывают на отрицательную автокорреляцию. В общем случае, значения близкие к 0 и 4 сигнализируют о сильной автокорреляции.

Тест Льюинга-Бокса:

$$Q = n(n+2) \sum_{k=1}^m \frac{\widehat{p}_k^2}{n-k}$$

$\widehat{p}_k$  - автокорреляции

Условия применения: Тест Льюинга-Бокса применяется для проверки автокорреляции остатков модели на нескольких лагах.

Интерпретация: Тест вычисляет статистику Льюинга-Бокса, которая сравнивает наблюдаемую автокорреляцию в остатках с автокорреляцией, ожидаемой при отсутствии автокорреляции (гипотеза о независимости). Если полученное значение статистики меньше критического значения, это говорит о том, что модель является адекватной и остатки не содержат значимой автокорреляции.

Оба теста используются для проверки гипотезы о независимости остатков модели. Если остатки модели обнаруживают значимую автокорреляцию, это указывает на неполноту модели и может потребоваться дальнейшее уточнение или модификация модели. Важно отметить, что эти тесты не ограничиваются только AR и MA моделями и могут применяться и к другим типам моделей временных рядов.

### 93. Линейная регрессия для стационарных рядов: Модель FDL.

Модель FDL (Fixed Distributed Lag) — это модель линейной регрессии, которая используется для анализа стационарных временных рядов с учетом лаговых зависимостей. В модели FDL зависимая переменная представляется как линейная комбинация текущих и отступающих значений независимых переменных.

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \dots + \beta_p X_{t-p} + \varepsilon_t$$

$Y_t$  - зависимая переменная в момент времени  $t$ ,

$X_t, X_{t-1}, \dots, X_{t-p}$  - независимые переменные (факторы),

соответствующие текущим и отступающим значениям,

$\beta_0, \beta_1, \dots, \beta_p$  - коэффициенты регрессии, отражающие влияние каждой независимой переменной на зависимую переменную,

$\varepsilon_t$  - случайная ошибка в момент времени  $t$ .

Модель FDL позволяет учесть лаговые зависимости в данных и выявить влияние прошлых значений независимых переменных на текущее значение зависимой переменной. Коэффициенты регрессии  $\beta_1, \beta_2, \dots, \beta_p$  отражают величину и статистическую значимость эффекта каждого отдельного лага на зависимую переменную.

Оценка коэффициентов в модели FDL обычно осуществляется с помощью метода наименьших квадратов (МНК) или других методов оценки параметров регрессии. При оценке модели FDL важно учитывать стационарность временного ряда и возможность автокорреляции остатков. Модель FDL может быть полезной для анализа и прогнозирования временных рядов, особенно в случаях, когда предшествующие значения независимых переменных имеют влияние на текущее значение зависимой переменной, и когда данные являются стационарными.

#### 94. Линейная регрессия для стационарных рядов. Модель ADL.

Использование линейной регрессии для стационарных временных рядов может быть расширено с помощью модели авторегрессии с распределенными лагами. Модель ADL позволяет учесть как лаги зависимой переменной, так и лаги независимых переменных при анализе временных рядов. Формально, модель ADL может быть записана следующим образом:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \dots + \gamma_q X_{t-q} + \varepsilon_t,$$

$Y_t$  - зависимая переменная в момент времени  $t$ ,

$Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$  - лаговые значения зависимой переменной,

$X_{t-1}, X_{t-2}, \dots, X_{t-q}$  - лаговые значения независимых переменных

$\beta_0, \beta_1, \dots, \beta_p$  - коэффициенты регрессии для лагов зависимой переменной,

$\gamma_1, \gamma_2, \dots, \gamma_q$  - коэффициенты регрессии для лагов независимых переменных,

$\varepsilon_t$  - случайная ошибка в момент времени  $t$ .

Модель ADL позволяет учесть как долгосрочные, так и краткосрочные зависимости между зависимой и независимыми переменными.

Коэффициенты регрессии  $\beta_1, \beta_2, \dots, \beta_p$  отражают долгосрочное влияние предыдущих значений зависимой переменной, а коэффициенты регрессии  $\gamma_1, \gamma_2, \dots, \gamma_q$  отражают краткосрочное влияние предыдущих значений независимых переменных.

Оценка коэффициентов в модели ADL также может быть выполнена с помощью метода наименьших квадратов (МНК) или других методов оценки параметров регрессии. При оценке модели ADL также важно учитывать стационарность временного ряда и возможность автокорреляции остатков.

Модель ADL может быть полезной для анализа и прогнозирования временных рядов, особенно в случаях, когда как предыдущие значения зависимой переменной, так и независимых переменных, имеют важное влияние на текущее значение зависимой переменной, и когда данные являются стационарными.

## 95. Понятие TS-ряда. Модель линейного тренда. Модель экспоненциального тренда.

Ряд  $x_t$  называется стационарным относительно тренда, если для него имеет место представление  $x_t = f(t) + v_t$  где  $f(t)$  – детерминированная функция (тренд или долгосрочная тенденция),  $v_t$  – стационарный ряд с нулевым средним. Следовательно,  $Ex_t = f(t)$  и  $Var(x_t) = Var(v_t) = const$

Замечание. В прикладных задачах типична ситуация, когда в модели ряд  $v_t$  автокоррелирован (следствие зависимости наблюдений)

### Модель линейного тренда задается

уравнением  $x_t = \beta_0 + \beta_1 t + u_t, t = 1, \dots, n$

Будем предполагать, что ошибки  $u_t$  удовлетворяют условиям теоремы Гаусса – Маркова, поэтому к модели линейного тренда применимы выводы стандартной линейной модели регрессии. В частности, среднее значение  $Ex_t$  линейно зависит от времени  $t$ :  $Ex_t = \beta_0 + \beta_1 t$

Коэффициент  $\beta_1$  имеет следующую интерпретацию: это есть среднее приращение временного ряда за один период времени.  $\Delta Ex_t = Ex_t - Ex_{t-1} = \beta_1$

Следовательно, с увеличением времени,

- при  $\beta_1 > 0$  во временном ряду есть «тенденция к возрастанию»,
- при  $\beta_1 < 0$  во временном ряду есть «тенденция к убыванию», причем средняя скорость изменения временного ряда за один период времени постоянна.

### Модель экспоненциального тренда

задается уравнением  $\ln(x_t) = \beta_0 + \beta_1 t + u_t$

Будем предполагать, что ошибки  $u_t$  удовлетворяют условиям теоремы Гаусса – Маркова. Тогда к модели линейного тренда применимы выводы стандартной линейной модели регрессии. В частности, среднее значение зависит от  $t$  экспоненциально.  $E\ln(x_t) = \beta_0 + \beta_1 t$

Для коэффициента  $\beta_1$  получаем следующую интерпретацию:

$$\Delta E\ln(x_t) = E\ln(x_t) - E\ln(x_{t-1}) = E\ln\left(\frac{x_t}{x_{t-1}}\right) = \beta_1$$

Следовательно, за один период времени (в среднем) значение  $x_t$  изменяется в  $\exp(\beta_1)$  раз. Если  $\beta_1$  мало, то  $\exp(\beta_1) \approx 1 + \beta_1$  и за один период времени в среднем значение  $x_t$  изменяется (в первом приближении) на  $\beta_1$ .

## 96. Нестационарные временные ряды: случайное блуждание, стохастический тренд, случайное блуждание со сносом.

модель AR (1)  $x_t = \phi x_{t-1} + u_t, u_t \sim WN(0, \sigma_u^2)$

А в случае  $\phi = 1$  ряд  $x_t$  называется случайным блужданием (random walk)

$$x_t = \phi x_{t-1} + u_t, u_t \sim WN(0, \sigma_u^2)$$

для случайного блуждания:

$$x_t = u_t + x_{t-1} = u_t + u_{t-1} + x_{t-2} = \dots = \sum_{s=1}^t u_s + x_0$$

Определение.

$\sum_{s=1}^t u_s$  называется стохастическим трендом ( $u_s$  – белый шум).

Особенность названия происходит от того, что локально ряд часто похож на детерминированный тренд (локально наблюдается тенденция к росту или к убыванию). Далее, если блуждание “выходит” из нуля ( $x_0 = 0$ )

$$x_t = \sum_{s=1}^t u_s, u_s \sim WN(0, \sigma_u^2)$$

то  $Ex_t = 0$

$Var(x_t) = t\sigma_u^2$  то есть разброс растет пропорционально  $\sqrt{t}$

$$cov(x_t, x_s) = \min(t, s)\sigma_u^2; corr(x_t, x_s) = \frac{\min(t, s)}{\sqrt{ts}}$$

Отсюда можно сделать вывод, что случайное блуждание нестационарно и есть сумма всех прошлых шоков.

модель AR(1):  $x_t = \mu + \phi x_{t-1} + u_t, u_t \sim WN(0, \sigma_u^2)$

При  $\phi = 1$  ряд  $x_t$  называется случайным блужданием со сносом (random walk with drift)

$$x_t = \mu + \phi x_{t-1} + u_t, u_t \sim WN(0, \sigma_u^2)$$

Для случайного блуждания со сносом, считая  $x_0 = 0$ , получаем

$$x_t = \mu + u_t + x_{t-1} = 2\mu + u_t + u_{t-1} + x_{t-2} = \dots = \mu t + \sum_{s=1}^t u_s$$

## 97. Дифференцирование ряда: определение, DS-ряды.

Операция дифференцирования или первая конечная разность ряда обозначается как  $\Delta x_t = x_t - x_{t-1} = (1 - L)x_t$  где  $L$  – лаговый оператор.

Идея состоит в том, чтобы вместо исходного ряда рассматривать его приращение за один период. Дифференцирование ряда позволяет перейти к стационарному ряду относительно приращений.

Определение. Ряд  $x_t$  называется интегрированным порядка  $k$ , если

1.  $x_t$  не TS-ряд и нестационарный;

2.  $k$  – минимальный порядок такой, что  $\Delta^k x_t$  – стационарный или TS-ряд.

Обозначение:  $x_t \sim I(k)$ .

Замечание. Если  $x_t$  – стационарный ряд, то формально  $x_t \sim I(0)$ .

Определение. Ряд  $x_t$  называется DS-рядом, если  $x_t \sim I(k)$  для некоторого  $k$ .

Запишем случайное блуждание как DS-ряд. Пусть  $x_t$  – случайное блуждание без сноса, тогда

$$x_t = x_{t-1} + u_t \Rightarrow \Delta x_t = x_t - x_{t-1} = u_t \sim WN(0, \sigma^2)$$

Следовательно,  $x_t \sim I(1)$  при этом, конечно,  $x_t$  не является TS-рядом.

Аналогично для случайного блуждания со сносом имеем

$$x_t = \mu + x_{t-1} + u_t \Rightarrow \Delta x_t = x_t - x_{t-1} = \mu + u_t \sim WN(\mu, \sigma^2)$$

Следовательно,  $x_t \sim I(1)$ .

## 98. Подход Бокса-Дженкинса.

Процедура Бокса-Дженкинса является методом идентификации, оценки и диагностики моделей ARIMA (авторегрессионной интегрированной скользящей средней) для анализа и прогнозирования временных рядов.

Она включает следующие шаги:

**Идентификация стационарности:** Первым шагом является проверка стационарности временного ряда. Стационарный ряд имеет постоянное среднее значение, дисперсию и автокорреляцию, что облегчает моделирование. Если ряд не является стационарным, применяется процедура дифференцирования, чтобы привести его к стационарному виду.

**Определение порядка дифференцирования:** если ряд нестационарный, проводится дифференцирование до достижения стационарности. Порядок дифференцирования (I) определяется как количество раз, которое необходимо применить операцию дифференцирования для достижения стационарности. Используются различные методы и критерии для определения порядка дифференцирования, такие как визуальный анализ, тест Дики-Фуллера и тест КПСС.

**Определение порядков авторегрессии (p) и скользящего среднего (q):** после достижения стационарности и определения порядка дифференцирования, необходимо определить порядки авторегрессии и скользящего среднего. Для этого анализируются автокорреляционная функция (ACF) и частная автокорреляционная функция (PACF) временного ряда. Порядок авторегрессии (p) определяется по последнему значимому лагу на PACF, а порядок скользящего среднего (q) - по последнему значимому лагу на ACF.

## 99. Модель ARIMA.

**ARIMA** — это модель, комбинирующая авторегрессионные и скользящие средние модели. ARIMA позволяет моделировать данные, не являющиеся стационарными, как это не требуется для AR- и MA-моделей. ARIMA включает три параметра: параметр авторегрессии (p), параметр скользящего среднего (q) и параметр интегрирования (d).

Пусть  $x_t \sim \text{ARIMA}(p, k, q)$ . Это означает, что  $\Delta^k x_t$  является стационарной ARMA(p, q), и может быть представлена  $(1 - \phi_1 L - \dots - \phi_p L^p) \Delta^k x_t = \mu + (1 + \theta_1 L + \dots + \theta_q L^q) u_t$  и все корни авторегрессионного многочлена  $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$  по модулю больше 1 (условие стационарности).

Тогда, используя  $\Delta^k x_t = (1 - L)^k x_t$  получаем  $\phi(L)(1 - L)^k x_t = \mu + \theta(L)u_t$

где  $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$

Введём многочлен  $\gamma(z) = \phi(z)(1 - z)^k = 1 - \gamma_1 z - \dots - \gamma_{p+k} z^{p+k}$

Он имеет единичный корень кратности k, а остальные p корней по модулю больше 1. 53 Тогда для ряда  $x_t \sim \text{ARIMA}(p, k, q)$

$$\gamma(L)x_t = \mu + \theta(L)u_t \Rightarrow x_t = \mu + \sum_{j=1}^q \gamma_j x_{t-j} + u_t + \sum_{s=1}^q \theta_s u_{t-s}$$

Таким образом, модель ARIMA (p, k, q) можно рассматривать как нестационарную ARMA(p+k, q), авторегрессионный многочлен которой имеет единичный корень кратности k, а остальные p корней по модулю больше единицы. Отсюда происходит и название, так называемых, рядов с единичным корнем.

Определение  $x_t \sim \text{ARIMA}(p, k, q)^2$  если

$$x_t \sim I(k)$$

$$\Delta^k x_t \sim \text{ARMA}(p, q)^3$$

$$\text{ARMA}(p, q) = \text{ARIMA}(p, 0, q)$$

## 100. Тест ADF на единичный корень.

единичный корень против стационарного ряда с нулевым средним (без константы) для модели  $x_t = \gamma_1 x_{t-1} + \dots + \gamma_p x_{t-p} + u_t$ ,  $u_t \sim WN(0, \sigma_u^2)$  тестируем гипотезу о том, что авторегрессионный многочлен имеет единичный корень кратности 1 (остальные корни по модулю больше 1)

$$\gamma(z) = 1 - \gamma_1 z - \dots - \gamma_p z^p$$

очевидно,  $z=1$  корень перепишем

$$\Delta x_t = \gamma_1 x_{t-1} + \theta_1 \Delta x_{t-1} + \dots + \theta_{p-1} \Delta x_{t-p+1} + u_t$$

$$\phi = \gamma_1 + \dots + \gamma_p - 1$$

$$\theta_1 = -\gamma_2 - \dots - \gamma_p$$

$$\theta_{p-1} = -\gamma_3 - \dots - \gamma_p$$

тогда алгоритм теста ADF с константой: оценить (OLS) преобразованное уравнение. тестируется статистика

$$ADF_t = DF_t = \frac{\hat{\phi}_{OLS}}{s.e.(\hat{\phi})}$$

критическое значение:  $t$  – специальное критическое значение теста Дики-Фуллера без константы. Вывод: отвергаем  $H_0$  (гипотезу единичного корня) при  $ADF_t < -\tau_{cr} < 0$ .

единичный корень против стационарного ряда в общем случае

$Ext = a$  и допущается что  $a \neq 0$

$$x_t = a + z_t \cdot E z_t \equiv 0$$

Для заданного  $p$  тестируем  $H_0: z \sim Arima(p-1, 1, 0)$   $H_1: z \sim Arima(p, 0)$

тестируем, что авторегрессионный член имеет единичный корень кратности 1 и  $\mu = 0$   
преобразуем

$$\Delta x_t = \mu + \phi x_{t-1} + \theta_1 \Delta x_{t-1} + \dots + \theta_{p-1} \Delta x_{t-p+1} + u_t$$

Тогда гипотеза единичного корня означает, что  $\mu = 0$  и  $\phi = 0$ , а стационарность будет при  $\phi < 0$ .

алгоритм аналогичен алгоритму тесту ADF с константой единичный корень против стационарного ряда относительно линейного тренда

Пусть  $x_t$  – ряд с линейным трендом  $x_t = \beta_0 + \beta_1 t + z_t$ ,  $E z_t \equiv 0$  Для заданного  $p$  тестируем

$H_0: z \sim Arima(p-1, 1, 0)$   $H_1: z \sim Arima(p, 0)$

тестируем, что авторегрессионный многочлен  $\gamma(z)$  имеет единичный корень кратности один (остальные корни по модулю больше 1) и  $\beta = 0$ . преобразуем

$$\Delta x_t = \mu + \beta t + \phi x_{t-1} + \theta_1 \Delta x_{t-1} + \dots + \theta_{p-1} \Delta x_{t-p+1} + u_t$$

Тогда гипотеза единичного корня означает, что  $\beta = 0$  и  $\phi = 0$ , а стационарность будет при  $\phi < 0$ .

алгоритм теста аналогичен тесту ADF с константой с учетом того, то оцениваем преобразованное уравнение при помощи МНК.

4. Тестирование правильности выбора спецификации: типичные ошибки спецификации модели, Критерий Акаике, Критерий Шварца. условия применения критериев.

Возможные ошибки спецификации модели:

1. Неправильный выбор вида уравнения регрессии
2. В уравнение регрессии включена лишняя (незначимая) переменная
3. В уравнении регрессии пропущена значимая переменная
- 4) Логическая модель не имеет смысла с точки зрения прикладной области

Статистика Акаике:

$$AIC = \ln\left(\frac{ESS_k}{n}\right) + \frac{2k}{n} + 1 + \ln(2\pi)$$

При увеличении объясняющих переменных первое слагаемое в правой части уменьшается, второе – увеличивается.

Среди нескольких альтернативных моделей (полной и редуцированной) предпочтение отдается модели, с наименьшим значением статистики AIC, в которой достигается определенный компромисс между величиной остаточной суммы квадратов и количеством объясняющих переменных. Особенности критерия: Штрафование числа параметров ограничивает значительный рост сложности модели, проверка критерия является трудоемкой операцией, может сравнивать модели только с выборками равного размера, порядок выбора моделей неважен.

Статистика Шварца:

$$SC = \ln\left(\frac{ESS_k}{n}\right) + \frac{k \ln(n)}{n} + 1 + \ln(2\pi)$$

При увеличении количества объясняющих переменных первое слагаемое в правой части уменьшается, а второе – увеличивается. Среди нескольких альтернативных моделей (полной и редуцированной) предпочтение отдается модели с наименьшим значением статистики Шварца. Используется при длинных выборках данных. Отличается следующими признаками: Из двух моделей предпочтительно выбрать с меньшим значением байесовского критерия. Байесовский критерий представляет собой возрастающую функцию от числа параметров модели и от остаточной суммы квадратов ошибок модели.

Изменение зависимых переменных и увеличение числа наблюдаемых увеличивает байесовский критерий, в то же время уменьшение критерия означает уменьшение размерности модели.

Используется при длинных выборках данных.

## 101. Модель ARCH.

Модель условной авторегрессионной гетероскедастичности – модель с такой формой гетероскедастичности, при которой последующие значения отклонений будут зависеть от величин предыдущих. В моделях типа ARCH используемые ряды предполагается стационарными. Простейшая форма модели

$$y_t = \beta'x_t + \varepsilon_t$$
$$\varepsilon_t = u_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}$$

$u_t$ - распределены по стандартному нормальному закону. Случайные остатки не зависят от значений факторов, входящих в модель регрессии. Для модели ARCH МНК дает лучшую линейную несмещенную оценку параметра  $\beta$ . Если расширить простейшую модель до  $\tau$ , получим ARCH( $\tau$ ).

$$y_t = \beta'x_t + \varepsilon_t$$
$$\sigma^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_q \varepsilon_{t-q}^2$$

Существуют также и модифицированный ARCH-M

$$y_t = \beta'x_t + \delta \sigma_t^2 + \varepsilon_t$$
$$\text{Var}[\varepsilon_t | \psi_t] = \text{ARCH}(q)$$

При выполнении основных предположений коэффициент  $\delta$  может трактоваться коэффициент относительной склонности к риску. Недостатки модели ARCH:

Как определить параметр  $q$ , определяющий количество лагов квадратов остатков модели?

Значение  $\tau$  количества лагов квадратов ошибок, которое необходимо для того, чтобы охватить все в зависимости условной дисперсии, может быть очень велико.

ограничение на неотрицательность коэффициентов может нарушаться.

## 102. Модель GARCH.

Модуль, в которой предполагается, что условная дисперсия будет зависеть также от собственных лагов.

Простейшая модель GARCH(1, 1) (поскольку использует первые лаги  $\mu_2$  и  $\sigma^2$ ).

$$\sigma^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

Основное преимущество модели GARCH заключается в том, что для спецификации моделей GARCH требуется меньше параметров. Модель большей степени будет удовлетворять условием неотрицательности.

Модель GARCH(1, 1) может быть расширена до модели GARCH(p, q)

$$\sigma^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \dots + \beta_p \sigma_{t-p}^2$$

Нецелесообразно использовать модели GARCH более высоких порядков, чем (1,1). Несмотря на то, что условная дисперсия модели GARCH изменяется со временем, безусловная дисперсия будет постоянной при  $\alpha + \beta < 1$

$$\text{var}(\varepsilon_t) = \frac{\alpha_0}{1 - (\alpha_1 + \beta)}$$

Если  $\alpha + \beta = 1$  модель становится IGARCH. В ней прогноз условной дисперсии равняется бесконечности. Для оценки GARCH моделей используется метод максимального правдоподобия.

### 103. Область применения панельных данных. Преимущества использования панельных данных.

Область применения – макроэкономика, микроэкономика, анализе финансово-хозяйственной деятельности, социологии и другие науки. Панельные данные чаще используются для анализа домохозяйств и индивидов.

Преимущества использования панельных данных.

1. Панельные данные позволяют учитывать индивидуальную неоднородность. Временные ряды или пространственные данные не всегда позволяют учесть неоднородность индивидов, фирм, регионов или стран, что может привести к смещенным оценкам. Панельные данные дают возможность избежать ошибки спецификации, возникающей из-за того, что существенные переменные не включены в модель.
2. Панельные данные содержат большое число наблюдений и тем самым предоставляют исследователю большее количество информации, им свойственна большая вариация и меньшая коллинеарность объясняющих переменных, они дают большее число степеней свободы и обеспечивают большую эффективность оценок.
3. Панельные данные предоставляют возможность изучать динамику изменений индивидуальных характеристик единиц совокупности. Панельные данные могут использоваться как для объяснения того, почему различные единицы совокупности ведут себя по-разному, так и для того, чтобы определить, почему конкретная единица совокупности ведет себя по-разному в различные периоды времени.
4. Панельные данные лучше способны идентифицировать и измерить эффекты, которые просто не определяемы только во временных рядах или только в пространственных данных. В качестве примера может выступать исследование того, происходит ли увеличение или уменьшение заработной платы за счет членства в профсоюзе. На этот вопрос лучше всего ответить, если мы наблюдаем переход работника с работы с профсоюзом на работу без профсоюза или наоборот, а это могут отразить только панельные данные. Рассматривая индивидуальную характеристику работника в качестве константы, можно будет определить, оказывает ли влияние членство в профсоюзе на зарплату и насколько. Подобный анализ может также использоваться для оценки других типов дифференциации заработной платы, например, для оценки премии, выплачиваемой за опасную или неприятную работу.
5. Панельные данные позволяют конструировать и тестировать более сложные поведенческие модели, чем пространственные данные и временные ряды в отдельности.
6. Панельные данные позволяют избежать смещения, связанного с агрегированием данных, так как панельные данные, собранные на микроуровне (по индивидам, фирмам или домашним хозяйствам), могут быть измерены более точно, чем аналогичные переменные, полученные на макроуровне. При этом во временных рядах рассматривается изменение во времени характеристик некоторой усредненной репрезентативной единицы совокупности, а в пространственных данных не учитываются ненаблюдаемые индивидуальные характеристики единиц совокупности.
7. Панельные данные макроуровня имеют более длинные временные ряды, и панельные тесты на единичный корень имеют стандартные асимптотические распределения в отличие от проблемы нестандартных распределений, типичной для теста на единичный корень в анализе временных рядов.

#### 104. Модели панельных данных и основные обозначения.

Пусть имеются данные  $y_{it}, x_{it}, i=\overline{1, N}, t=\overline{1, T}$  Здесь  $N$  – количество субъектов, а  $T$  – число последовательных моментов времени. Требуется оценить модель линейной связи между переменными  $Y$  и  $X$ . В общем случае  $X$  является вектором конечной размерности  $k$  (может существовать  $p$  независимых факторов).

Рассмотрим сбалансированные панели, где для каждой пространственной единицы имеется одинаковое число наблюдений по всем периодам времени. Тогда общее число наблюдений будет  $N \cdot T$ . При  $N=1$  и достаточно большом  $T$  получаются временные ряды, а при  $T=1$  и достаточно большом  $N$  получаются пространственные данные. Метод оценивания панельных данных относится к случаю, когда  $N > 1$  и  $T > 1$ .

Будем рассматривать панельные данные с короткими временными рядами, где  $N$  намного больше  $T$ , что очень часто встречается на практике, когда число наблюдаемых единиц достаточно велико (может достигать нескольких сотен или тысяч), а число моментов наблюдения ограничено.

Для  $i$ -й единицы совокупности данные можно представить в виде

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{iT} \end{bmatrix}, X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{iT} \end{bmatrix}, \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \\ \varepsilon_{iT} \end{bmatrix}$$

### 105. Модель пула (Pool model).

При отсутствии значимых различий (неоднородности) между пространственными объектами выборки, возможно построение регрессии по объединенной выборке (pooled regression) – пула. Это модель сквозной регрессии:

$$y_{it} = X_{it}b + a_i + \varepsilon_{it}$$

с остатками  $\varepsilon_{it}$ , удовлетворяющими требованиям МНК. В этом случае мы имеем дело с обычной линейной регрессией с  $N \times T$  наблюдениями, удовлетворяющей предположениям классической нормальной линейной модели. Для получения эффективных оценок вектора коэффициентов достаточно использовать обычный метод наименьших квадратов (OLS). Полученные при этом оценки  $b$  и  $a$  являются наилучшими линейными несмещенными оценками (BLUE – best linear unbiased estimate) вектора  $\beta$ . При соответствующих предположениях о поведении значений объясняющих переменных, когда  $N \rightarrow \infty$  или/и  $T \rightarrow \infty$ , эта оценка является также и состоятельной оценкой этого вектора.

## 106. Модель регрессии с фиксированным эффектом (fixed effect model)

$$y_{it} = X_{it}b + a_i + \varepsilon_{it} \quad (1)$$

Свободный член  $a_i$  принимает различные значения для каждого объекта выборки.

Смысл его в том, чтобы отразить влияние пропущенных или ненаблюдаемых переменных, характеризующих индивидуальные особенности исследуемых объектов не меняющиеся со временем. Термин "фиксированные эффекты" означает, что константа в уравнении регрессии может различаться между объектами, но для каждого конкретного объекта константа является постоянной во времени, т. е. не изменяется с течением времени  $t$ .

$$\text{Уравнение модели в матричной записи } y = X \cdot b + Z \cdot A + \varepsilon \quad (2)$$

Размерности матриц:  $y$  ( $NT \times 1$ ),  $X$  ( $NT \times k$ ),  $b$  ( $k \times 1$ ),  $Z$  ( $NT \times N$ ),  $A$  ( $N \times 1$ ),  $\varepsilon$  ( $NT \times 1$ )

Оценка коэффициентов  $\beta$ , где  $Z$  – блочно-диагональная матрица фиктивных переменных. Считаются через МНК.

$$\beta_{LSDV} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'Y \\ Z'Y \end{pmatrix}$$

Считаются через обычный МНК

Те же самые значения оценок параметров  $\beta$  и  $a$  можно получить иным способом. Пусть

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}, \bar{X}_i = \frac{1}{T} \sum_{t=1}^T X_{it}, \bar{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \varepsilon_{it}$$

где под  $X$  понимается вектор из всех факторов  $x_k$ . То есть надо рассчитать средние по времени для каждого из факторов. Усредняя по времени обе части уравнения (1), получим

$$\bar{y}_i = \bar{X}_i b + a_i + \bar{\varepsilon}_i \quad (3), i = \overline{1, N}$$

Из уравнений (1) и (3) получаем

$$y_{it}^* = X_{it}^* b + \varepsilon_{it}^*$$

Где  $y_{it}^* = y_{it} - \bar{y}_i$ ,  $X_{it}^* = X_{it} - \bar{X}_i$ ,  $\varepsilon_{it}^* = \varepsilon_{it} - \bar{\varepsilon}_i$

В результате мы получили модель, скорректированную на индивидуальные средние. Свободные члены оказались исключенными из уравнения. По полученному уравнению можно рассчитать параметр  $b$ . Эта оценка называется «внутригрупповой» («within-group» estimate), имея в виду, что она строится только на основании отклонений значений переменных от их средних по времени и тем самым принимает во внимание только изменчивость в пределах каждого субъекта, не обращая внимание на изменчивость между субъектами. Получив параметр  $b$ , оценку  $a_i$  можно получить  $a_i = \bar{y}_i - \bar{X}_i b, i = \overline{1, N} \quad (4)$

Полученные в итоге этих двух методов оценки  $a$  и  $b$  численно совпадают, но стандартные ошибки оценок  $\beta$  отличаются в этих двух моделях. При «within»-оценке стандартная ошибка меньше. А стандартные ошибки оценок  $a_i$ , получаемые по (4), нельзя вычислять по формулам для стандартных ошибок оценок наименьших квадратов. Таким образом,  $\beta$  является состоятельной оценкой и когда  $N \rightarrow \infty$  и когда  $T \rightarrow \infty$ , в то время как  $a_i$  состоятельна только, когда  $T \rightarrow \infty$ . Последнее есть следствие того, что оценивание каждого  $a_i$  производится фактически лишь по  $T$  наблюдениям, так что при фиксированном  $T$  с ростом  $N$  происходит лишь увеличение количества параметров  $a_i$ , но это не приводит к возрастанию точности оценивания каждого конкретного  $a_i$

## 107. Модель регрессии со случайным эффектом (random effect model).

В случае отсутствия корреляции между индивидуальными эффектами и регрессорами, индивидуальные эффекты можно рассматривать как одну из составляющих ошибки.

Модель со случайными эффектами имеет вид

$$y_{it} = x_{it}'\beta + \alpha + u_i + \varepsilon_{it}$$

$u_i + \varepsilon_{it}$  - составная ошибка регрессии, которая содержит 2 компоненты индивидуальную компоненту  $u_i$  и остаточный член  $\varepsilon_{it}$ . Компонента  $u_i$  - с индивидуальной ошибки, которая является постоянной во времени для объекта.

В матричной форме

$$y = Xb + u$$

Матожидание составной ошибки равно 0. Ошибки в модели со случайными эффектами являются гетероскедастичными, и для получения эффективных оценок параметров к необходимо применять обобщенный метод наименьших квадратов. Обобщенная оценка наименьших квадратов учитывает и внутригрупповую, и межгрупповую изменчивость.

$$b = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

Для преобразования компонент используются операторы Between (усреднения по времени для отдельного объекта) или Within (вычисления отклонения от среднего по времени для отдельного объекта).

Оценить дисперсию  $\sigma_\alpha^2$  случайных эффектов можно заметить, что при оценивании модели

$$\bar{y}_i = \mu + \bar{X}_i b + \bar{u}_i$$

приводящей к межгрупповой оценке  $\bar{\beta}_B$  дисперсия остатка для  $i$ -ой группы равна  $D(\bar{u}_{it}) = \sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{T}$  откуда находим  $\sigma_\alpha^2$ .

Алгоритм нахождения  $\beta$

Находим  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$ ,  $\bar{X}_i = \frac{1}{T} \sum_{t=1}^T X_{it}$

по полученным  $\bar{y}_i$  и  $\bar{X}_i$  строим регрессию и находим остаточную дисперсию  $D(\bar{u}_{it})$

Находим  $\sigma_\alpha^2 = D(\bar{u}_{it}) - \frac{\sigma_\varepsilon^2}{T}$   $\sigma_\varepsilon^2$  - остаточная дисперсия из модели с фиксированными эффектами

Находим  $\theta = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}$  и  $1 - \theta$

Переходим к новым переменным  $\tilde{X}_{it} = X_{it} - (1 - \theta)\bar{X}_i$ ,  $\tilde{y} = y_{it} - (1 - \theta)\bar{y}_i$

Строим по ним регрессии получаем оценки  $\beta$  коэффициенты в модели со случайными эффектами

Для сравнения модели с со случайными эффектами со сквозной регрессией используется тест Бройша – Пагана

## 108 Тест Бройша-Пагана для панельных данных 108

Для сравнения модели со случайными эффектами со сквозной регрессией (моделью пула) используется тест Бройша-Пагана. Это критерий для проверки в рамках RE-модели (со стандартными предположениями) гипотезы о равенстве межгрупповой дисперсии ошибок нулю:

$$H_0: \sigma_\alpha^2 = 0 \text{ — сведение к модели пула}$$

Статистика критерия Бройша-Пагана

$$BP = \frac{NT}{2(T-1)} \left( \frac{\sum_{i=1}^N (\sum_{t=1}^T \varepsilon_{it})^2}{\sum_{i=1}^N \sum_{t=1}^T \varepsilon_{it}^2} - 1 \right)^2$$

Здесь остатки берутся из модели пула.

Статистика критерия Бройша-Пагана распределена по  $\chi^2(1)$ .

Соответственно, гипотеза  $H_0$  отвергается в пользу модели со случайными эффектами, если наблюдаемое значение статистики ВР превышает критическое значение, рассчитанное по распределению  $\chi^2(1)$ .

## 109 Тест Хаусмана для панельных данных 109

Для сравнения модели со случайными эффектами с моделью с фиксированными эффектами используется тест Хаусмана. Модель со случайным эффектом имеет место только в случае некоррелированности случайного эффекта с регрессорами. Это требование часто бывает нарушено. Как было показано Мундлаком, учет подобной корреляции приводит к регрессии, в которой МНК-оценки коэффициентов наклона совпадают с оценками «within». В тесте проверяются следующие гипотезы:

$$H_0: cov(a_i, X_{it}) = 0$$

$$H_1: cov(a_i, X_{it}) \neq 0$$

Можно проводить сразу несколько версий теста Хаусмана, вводя

$$q_1 = \hat{b}_W - \hat{b}_{\text{ОМНК}} \quad q_2 = \hat{b}_W - \hat{b}_B \quad q_3 = \hat{b}_{\text{ОМНК}} - \hat{b}_B$$

Для всех трех вариантов проверяется критерий

$$m = q' \Omega^{-1} q$$

Который распределен по  $\chi^2(k)$ .

Если остаемся в области  $H_0$ , следует пользоваться моделью RE со случайными эффектами. Если попадаем в критическую область, следует пользоваться моделью FE с фиксированными эффектами.

## 110 Тест Лагранжа для панельных данных 110

Обычная модель против модели со случайным эффектом. В этом случае требуется в модели со случайным эффектом тестировать гипотезу  $H_0: \sigma_u^2 = 0$  Бреуш и Паган предложили тест множителей Лагранжа, основанный на следующей статистике

$$LM = \frac{nT}{2(T-1)} \left( \frac{\sum_{i=1}^n (\sum_{t=1}^T \varepsilon_{it})^2}{\sum_{i=1}^n \sum_{t=1}^T \varepsilon_{it}^2} - 1 \right)^2$$

Где  $\varepsilon_{it}$ -остатки в обычной регрессии. При гипотезе  $H_0$  величина LM имеет хи-квадрат распределение с одной степенью свободы. Используя матрицу D фиктивных переменных, можно более компактно представить статистику LM:

$$LM = \frac{nT}{2(T-1)} \left( \frac{e' D D' e}{e' e} - 1 \right)^2$$

Как обычно, если  $LM > \chi_{\alpha}^2(1)$ , то гипотеза  $H_0$  отвергается при уровне значимости  $\alpha$ , где  $\chi_{\alpha}^2(1)$  –  $\alpha$ -процентная точка распределения хи-квадрат с одной степенью свободы.

## 111 Вычисление значения оценок параметров $\beta$ и $a$ в модели с фиксированным эффектом 111

Уравнение модели в матричной записи

$$y = X * b + Z * A + \varepsilon \quad (3)$$

Размерности матриц:

$$y(NT \times 1), X(NT \times k), b(k \times 1), Z(NT \times N), A(N \times 1), \varepsilon(NT \times 1)$$

$A$  – вектор констант, соответствующих детерминированным индивидуальным эффектам, а  $Z$ -блочная диагональная матрица фиктивных переменных

$$\begin{bmatrix} y_1 \\ (T, 1) \\ \dots \\ y_N \\ (T, 1) \end{bmatrix} = \begin{bmatrix} X_1 \\ (T, K) \\ \dots \\ X_N \\ (T, K) \end{bmatrix} * b(K, 1) + \begin{bmatrix} i_T & 0 & \dots & 0 \\ 0 & \vec{1}_T & & \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \vec{1}_T \end{bmatrix} * \begin{bmatrix} a_1 \\ \dots \\ a_i \\ \dots \\ a_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_N \end{bmatrix}$$

В этом случае оценка коэффициентов  $\beta$

$$\widehat{\beta}_{LSDV} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}^{-1} \begin{pmatrix} X'Y \\ Z'Y \end{pmatrix}$$

Считаются через обычный МНК.

Эта модель является довольно гибкой, так как, в отличие от предыдущей модели, она позволяет учитывать индивидуальную гетерогенность объектов. Однако, за эту гибкость часто приходится расплачиваться потерей значимости оценок (из-за увеличения их стандартных ошибок), так как приходится оценивать  $N$  лишних параметров. Если количество субъектов анализа  $N$  велико, необходимость обращать матрицу высокой размерности  $(N+K)$  вызывает вычислительные трудности. Интересно, что численно те же самые значения оценок параметров  $\beta$  и  $a$  можно получить иным способом. Пусть

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}, \bar{X}_i = \frac{1}{T} \sum_{t=1}^T X_{it}, \bar{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \varepsilon_{it},$$

Где под  $X$  понимается вектор из всех факторов  $x_k$ . То есть надо рассчитать средние по времени для каждого из факторов. Усредняя по времени обе части уравнения(2), получим  $\bar{y}_i = \bar{X}_i b + a_i + \bar{\varepsilon}_i$ ,  $i = 1, N$  (4)

Из уравнений (2) и (4) получаем  $y_{it} = X_{it} b + \varepsilon_{it}$ ,

где  $\hat{y}_{it} = y_{it} - \bar{y}_i$ ,  $\hat{X}_{it} = X_{it} - \bar{X}_i$ ,  $\hat{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_i$

В результате мы получили модель, скорректированную на индивидуальные средние. Свободные члены оказались исключенными из уравнения. По полученному уравнению можно рассчитать параметр  $b$ . Эта оценка называется «внутригрупповой», имея в виду, что она строится только на основании отклонений значений переменных от их средних по времени и тем самым принимает во внимание только изменчивость в пределах каждого субъекта, не обращая внимание на изменчивость между субъектами. Впрочем, в последнее время в эконометрической литературе чаще стали говорить об указанной оценке просто как о «within» - оценке.

Получив параметр  $b$ , оценку  $a_i$  можно получить

$$a_i = \bar{y}_i - \bar{X}_i b, i = 1, N$$

Полученные в итоге этих двух методов оценки  $a$  и  $b$  численно совпадают, но стандартные ошибки оценок  $\beta$  отличаются в этих двух моделях. При «within»-оценке стандартная ошибка меньше. А стандартные ошибки оценок  $a_i$ , получаемые по (6), нельзя вычислять по формулам для стандартных ошибок оценок наименьших квадратов.

Таким образом,  $\beta$  является состоятельной оценкой и когда  $N \rightarrow \infty$  и когда  $T \rightarrow \infty$ , в то время как  $a_i$  состоятельна только, когда  $T \rightarrow \infty$ . Последнее есть следствие того, что оценивание каждого  $a_i$  производится фактически лишь по  $T$  наблюдениям, так что при фиксированном  $T$  с ростом  $N$  происходит лишь увеличение количества параметров  $a_i$ , но это не приводит к возрастанию точности оценивания каждого конкретного  $a_i$

## 112 Отражение пространственных эффектов. Бинарная матрица граничных соседей. Приведите пример 112

Методы пространственной эконометрики предполагают исследование зависимостей одной территориальной системы от всех остальных.

При этом, ближайшие территориальные системы больше связаны друг с другом, чем расположенные на значительном расстоянии

Для исследования взаимосвязей между территориальными системами и используются матрицы пространственных весов.

Данные матрицы в табличной форме отражают расстояния между различными объектами в пространстве

Строки матрицы содержат веса для объекта в пространстве, на который оказывают влияние соседние объекты

Главная диагональ матрицы состоит из нулей, и таким образом исключается влияние объекта самого на себя

Зачастую весовая матрица нормализуется по строкам (сумма весов по каждой строчке матрицы равняется единице). Такая нормализация позволяет учитывать относительные, а не абсолютные расстояние между объектами.

Бинарная матрица граничных соседей

$$W_{ij} = \begin{cases} 0, & \text{если } i = j \\ 1, & \text{если } j \text{ граничит с } i, \\ 0, & \text{не граничит} \end{cases}$$

Это наиболее простой способ учета пространственных взаимосвязей

Согласно данной матрице на исследуемые объекты оказывают влияние только те соседи, которые граничат с ними

Из-за бинарности матрицы при нормализации ее значений получается, что на территорию оказывается влияние соседних территорий с одними и теми же пространственными весами.

Данную матрицу можно использовать, когда исследуемые территориальные системы достаточно однородны по площади

### 113 Отражение пространственных эффектов. Бинарная матрица ближайших соседей. Приведите пример 113

Методы пространственной эконометрики предполагают исследование зависимостей одной территориальной системы от всех остальных.

При этом, ближайшие территориальные системы больше связаны друг с другом, чем расположенные на значительном расстоянии.

Для исследования взаимосвязей между территориальными системами и используются матрицы пространственных весов.

Данные матрицы в табличной форме отражают расстояния между различными объектами в пространстве

Строки матрицы содержат веса для объекта в пространстве, на которые оказывают влияние соседние объекты

Главная диагональ матрицы состоит из нулей, и таким образом исключается влияние объекта самого на себя

Зачастую, весовая матрица нормализуется по строкам (сумма весов по каждой строчке матрицы равняется единице). Такая нормализация позволяет учитывать, относительные, а не абсолютные расстояния между объектами.

Бинарная матрица ближайших соседей

$$W_{ij}(k) = \begin{cases} 0, & \text{если } i = j \\ 1, & \text{если } d_{ij} \leq d_i(k), \\ 0, & \text{если } d_{ij} > d_i(k) \end{cases}$$

$d_i(k)$  – минимальное расстояние k-го порядка между объектами  $i$  и  $j$

В этом случае число соседей для каждого объекта будет равно  $k$ .

Рассчитываются расстояния от данного объекта до всех имеющихся объектов.

Затем берется  $k$  минимальных расстояний

И  $k$ -е расстояние для данного объекта, для получения устойчивых результатов учитываются 10-25 ближайших соседей

## 114 Отражение пространственных эффектов. Матрица расстояний.

### Приведите пример 114

Методы пространственной эконометрики предполагают исследование зависимостей одной территориальной системы от всех остальных.

При этом, ближайшие территориальные системы больше связаны друг с другом, чем расположенные на значительном расстоянии

Для исследования взаимосвязей между территориальными системами и используются матрицы пространственных весов.

Данные матрицы в табличной форме отражают расстояния между различными объектами в пространстве.

Строки матрицы содержат веса для объекта в пространстве, на который оказывают влияние соседние объекты.

Главная диагональ матрицы состоит из нулей, и таким образом исключается влияние объекта самого на себя

Зачастую, весовая матрица нормализуется по строкам (сумма весов по каждой строчке матрицы равняется единице). Такая нормализация позволяет учитывать относительные, а не абсолютные расстояния между объектами.

Матрица расстояний

$$W_{ij}(k) = \begin{cases} 0, & \text{если } i = j \\ 1/d_{ij}^k, & \text{если } d_{ij} \leq D(q), \\ 0, & \text{если } d_{ij} > D(q) \end{cases}$$

$D(q)$  – квартили расстояний

$d_{ij}$  = расстояние между объектами

Матрица расстояний является аналогом гравитационной модели – притяжение объектов обратно пропорционально квадрату расстояния между ними.

Поэтому, чем дальше располагаются объекты друг от друга, тем меньше они взаимодействуют.

Если  $q < 4$ , то соответствующий квартиль расстояния  $D(q)$  является максимальным расстоянием, дальше которого взаимодействие между объектами является несущественным.

Если  $q = 4$ , то происходит учет всех расстояний (в весовой матрице нулевыми будут только элементы главной диагонали).

### 115. Отражение пространственных эффектов. Матрица расстояний с учетом размера объекта. Приведите пример.

Отражение пространственных эффектов в анализе данных может включать учет расстояний между объектами с учетом их размера. Это может быть полезно при изучении пространственной автокорреляции или при моделировании зависимостей, которые могут быть связаны с расстоянием и размером объектов.

Примером такого подхода может быть анализ распределения заболеваемости определенным инфекционным заболеванием в городе.

Предположим, что у вас есть данные о количестве заболевших и географических координатах различных районов города. Помимо этого, у вас также есть информация о размере каждого района, которая может быть представлена, например, площадью или населением.

Для учета размера объекта и его влияния на расстояние можно построить матрицу расстояний с учетом размера. В этой матрице расстояний каждое расстояние между объектами масштабируется на основе их размера. Такой подход позволяет учесть, что большие объекты могут оказывать большее влияние на расстояние и, следовательно, на пространственную структуру данных.

Применительно к нашему примеру, при построении матрицы расстояний с учетом размера объекта, расстояния между районами будут отмасштабированы в соответствии с их размером. Например, если два района имеют одинаковое географическое расстояние, но один из них является значительно больше по размеру, то его вес в матрице расстояний будет выше, чтобы отразить большее влияние этого района на общую пространственную структуру данных.

Такой подход может быть полезным для более точного моделирования пространственной структуры данных, особенно когда размер объектов может быть важным фактором в анализе.

## 116. Алгоритм построения матрицы пространственных весов.

Приведите пример.

Для учета размера объекта и его влияния на расстояние можно построить матрицу расстояний с учетом размера. В этой матрице расстояний каждое расстояние между объектами масштабируется на основе их размера. Такой подход позволяет учесть, что большие объекты могут оказывать большее влияние на расстояние и, следовательно, на пространственную структуру данных.

Алгоритм построения матрицы пространственных весов

1. Формирование матрицы расстояний административными центрами субъектов РФ ( $X_{ij}$ )

по линейным расстояниям;

по автомобильным дорогам, между железнодорожным магистралям, авиационным, речным сообщениям

по смежным границам (с использованием бинарных переменных 1 и 0).

2. Стандартизация расстояний в матрице по строкам

3. Преобразование матрицы расстояний в относительную  $V_{ij} = \frac{1}{X_{ij}}$

4. Формирование матрицы стандартизированных дистанций

между территориями  $W_{ij} = \frac{V_{ij}}{\sum V_{ij}}$

## 117. Пространственная автокорреляция по методологии А. Гетиса и Дж. Орда. Недостатки методологии.

Методология А. Гетиса и Дж. Орда, также известная как метод пространственной автокорреляции Гетиса-Орда, используется для анализа пространственной зависимости в данных. Этот метод позволяет оценить степень пространственной автокорреляции, то есть схожести значений переменных в пространстве.

$$G = \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} x_i x_j}{\sum_{i=1}^N \sum_{j=1}^N x_i x_j}$$
$$E(G) = \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}}{N(N-1)}$$
$$Z_G = \frac{G - E(G)}{\sqrt{V(G)}}$$

$G > E(G)$  - Наблюдается пространственная кластеризация объектов с высокими значениями

$G < E(G)$  - Наблюдается пространственная кластеризация объектов с низкими значениями

Рост  $Z$  - Повышение интенсивности пространственной кластеризации

Показатель Getis-Ord ( $G$ ) используется когда:

данные распределены достаточно равномерно,

необходимо найти неожиданные всплески высоких значений в пространстве

Несмотря на преимущества методологии Гетиса-Орда в анализе

пространственной зависимости, есть и некоторые недостатки:

Чувствительность к выбору весов: Метод Гетиса-Орда требует выбора специфических весов, которые определяют степень взаимосвязи между объектами. Выбор весов может оказывать значительное влияние на результаты и интерпретацию пространственной автокорреляции, и неправильный выбор весов может привести к искаженным выводам.

Зависимость от единичного масштаба: Метод Гетиса-Орда чувствителен к масштабу данных, поскольку использует суммы значений переменных. Это означает, что результаты могут изменяться при изменении масштаба переменных, что затрудняет сравнение результатов между разными наборами данных.

Игнорирование пространственной структуры: Метод Гетиса-Орда предполагает постоянную пространственную структуру зависимой переменной, не учитывая ее изменения или наличие пространственных кластеров. Это может привести к упущению важной информации о пространственных зависимостях в данных.

Ограниченная интерпретация: Интерпретация индекса пространственной автокорреляции Moran's  $I$  может быть сложной. Он указывает только на наличие пространственной автокорреляции, но не дает непосредственной информации о причинах или механизмах этой автокорреляции.

## 118. Пространственная автокорреляция по методологии Роберта Джири.

Мера Джири  $C$  является более чувствительной к локальной пространственной автокорреляции, чем Индекс П. Морана.

$$C = \frac{(n-1)}{2S_0} \frac{\sum_i^n \sum_j^n w_{ij} (Y_i - Y_j)^2}{\sum_i^n (Y_i - \bar{Y})^2}, S_0 = \sum_i^n \sum_j^n w_{ij}$$

Если  $C=1$  – Пространственная корреляция отсутствует

Если  $0 \leq C \leq 1$  Положительная пространственная корреляция

Если  $-1 \leq C \leq 0$  Отрицательная пространственная корреляция

**Оценка значимости статистики Джири**

$$Z = \frac{(C - E(C))}{SD(C)}$$

Величина  $Z$  определяет:

на какое количество стандартных отклонений фактическое значение статистики Р. Джири удалено от ожидаемого среднего значения ( $E$ )

чем сильнее оно удалено от распределения случайно, тем менее вероятно, что фактическое

при  $Z = E(C)$  - значения наблюдений в соседних территориях расположены случайным образом

## 119. Пространственная автокорреляция по методологии Морана П.

$$I_G = \frac{N}{\sum \sum W_{ij}} * \frac{\sum \sum W_{ij}(x_i - \mu)(x_j - \mu)}{\sum (x_i - \mu)^2}$$

где, N - число регионов;

$W_{ij}$  – элемент матрицы пространственных весов для регионов i и j;

$\mu$  - среднее значение показателя;

X - анализируемый показатель.

$E(I) = -\frac{1}{N-1}$  – среднее значение индекса

При  $I_G > E(I)$  наблюдается положительная пространственная автокорреляция (значения наблюдений в соседних территориях являются подобными)

При  $I_G < E(I)$  — наблюдается отрицательная автокорреляция (значения наблюдений в соседних территориях отличаются).

При  $I_G = E(I)$  — значения наблюдений в соседних территориях расположены случайным образом

Величина Z определяет:

$$z = \frac{I - E(I)}{\sqrt{var(I)}}$$

на какое количество стандартных отклонений фактическое значение индекса Морана удалено от ожидаемого среднего значения (E) чем сильнее оно удалено распределение случайно.

## 120. Пространственная кластеризация территорий. Локальный индекс автокорреляции П. Морана (I<sub>li</sub>)

Локальные индексы автокорреляции П. Морана позволяют:

Установить полюса роста пространственных кластеров

Оценить силу взаимовлияния полюсов роста на другие территории

Оценить направление пространственной корреляции (прямая / обратная)

$$I_{Li} = N * \frac{(x_i - \mu) * \sum w_{ij}(x_j - \mu)}{\sum (x_i - \mu)^2}$$

При  $I_{Li} < 0$  наблюдается отрицательная автокорреляция для территории  $i$ , данная территория существенно отличается по исследуемому показателю от соседних территорий (outlier)

При  $I_{Li} > 0$  — автокорреляция положительная, данная территория по исследуемому показателю подобна соседним территориям (cluster)

При  $|I_{Li}| > |I_{Lj}|$  — подобие/различие территории  $i$  с окружающими ее соседними территориями является большим, чем в случае территории  $j$  и ее соседей.

## 121. Матрица взаимовлияния Л. Анселина (LISA).

Матрица взаимовлияния позволяет:

оценить тесноту взаимосвязи между исследуемыми объектами в пространстве

выявить направление данных взаимосвязей (прямые и обратные)

$$LISA_{ij} = Z_i \cdot Z_j \cdot W_{ij}$$

LISA - индекс локальной автокорреляции между двумя регионами;

W - элемент матрицы пространственных весов для регионов i и j;

Z - стандартизированные значения показателя одного региона;

Z - стандартизированные значения показателя одного региона

$$Z_i = \frac{(x_i - \mu)}{\sqrt{\frac{\sum (x_i - \bar{x}_l)^2}{n}}}$$

Выделение в матрице значений, превышающих среднее значение локального индекса автокорреляции, позволит:

выявить зоны взаимовлияния полюсов роста,

установить территории, получающие импульс от их развития или способствующие их развитию