

**Основы построения
телекоммуникационных
систем и сетей**

**Лекция №15
«Методы прогнозирования»**

профессор Соколов Н.А.

Анализ временных рядов (1)

Можно выделить три группы методов прогнозирования временных рядов: наивные, усредняющие и сглаживающие. Наивный метод основан на предположении, что будущее лучше всего характеризуется последними изменениями. Метод усреднения позволяет разработать прогноз, основываясь на среднем значении прошлых наблюдений. Методы сглаживания базируются на усреднении накопленных данных при помощи набора весовых коэффициентов.

Корректный подход к оценке метода прогнозирования включает несколько этапов. Следует выделить пять важных этапов:

- тщательное изучение природы исследуемого объекта или процесса для выбора адекватного метода прогнозирования;
- выделение двух групп среди доступных данных – для разработки прогнозов и для проверки полученных результатов;
- уточнение исходных данных с целью обнаружения ошибок;
- разработка прогнозов и оценка достоверности полученных результатов;
- использование (интерпретация) полученных результатов и выполнение, при необходимости, уточнения и дополнения прогнозов.

Анализ временных рядов (2)

Если исходные данные представлены небольшим набором сведений, то наивный метод считается единственно возможным способом прогнозирования. При наивном прогнозировании считается, что последний период лучше всего предсказывает будущее. Самая простая модель наивного прогнозирования основана на следующем соотношении:

$$\hat{Y}_{t+1} = Y_t.$$

Величина \hat{Y}_{t+1} представляет собой прогноз, сделанный в момент времени t (начальное предсказание) для момента времени $t+1$. Наивным прогнозом для каждого периода времени является непосредственно предшествующее ему наблюдение. Текущему значению величины в ряду присваивается стопроцентный вес. Поэтому наивный прогноз часто называют "прогнозом без изменений". Фактически, при наивном прогнозировании отбрасываются все старые наблюдения.

В данном случае важно оценить случайные флуктуации исследуемого объекта или процесса. Ошибка прогноза r_{t+1} рассчитывается по очевидной формуле:

$$r_{t+1} = Y_{t+1} - \hat{Y}_{t+1}.$$

Анализ временных рядов (3)

Анализ статистики величин r_j позволяет сделать вывод о возможности использования наивного метода для исследуемого объекта или процесса. Очевидно, что при стремлении суммы величин r_j к нулю применение наивного прогнозирования может быть оправданно. С другой стороны, условие близости к нулю суммарной ошибки нельзя считать достаточным. Необходимо также чтобы максимальное значение r_j не превысило заранее выбранный порог ошибки. Если исследование ряда r_j указывает на наличие тренда для исследуемого объекта или процесса, то наивный метод не годится для разработки прогностических оценок.

Некоторые прогнозы приходится обновлять регулярно (ежедневно, еженедельно и так далее). Количество хранимой информации резко возрастает. Кроме того, "старая информация" может оказаться бесполезной. В подобных случаях прибегают к усреднению или к сглаживанию. При использовании подобных методик взвешенное усреднение данных, связанных с прошлыми наблюдениями, применяется для сглаживания случайных флуктуаций. Такой подход основан на том, что флуктуации являются случайными отклонениями от некоторой гладкой кривой. Эту кривую и следует определить.

Анализ временных рядов (4)

Методы усреднения, в свою очередь, делятся на несколько видов. Используют простые средние, скользящие средние и двойные скользящие средние.

В первом методе в качестве начальных данных используется среднее значение исследуемой величины в момент времени t , а в качестве тестовой части – остальные.

Прогноз \hat{Y}_{t+1} осуществляется на основании усреднения начальных данных:

$$\hat{Y}_{t+1} = \frac{1}{t} \sum_{j=1}^t Y_j .$$

Когда становится известным новое наблюдение Y_{j+1} , оно используется для разработки прогноза \hat{Y}_{t+2} . Схожесть с методом наивного прогнозирования наблюдается в случае применения искусственного приема, когда сохраняются только "свежие" прогнозы:

$$\hat{Y}_{t+2} = \frac{t\hat{Y}_{t+1} + Y_t}{t+1} .$$

Метод простых средних приемлем в тех случаях, когда исследуемые объекты или процессы стабильны. Характерный пример такой ситуации – объем продаж товара, зависящий от усилий торговой сети (без сезонных колебаний).

Линейная регрессия (1)

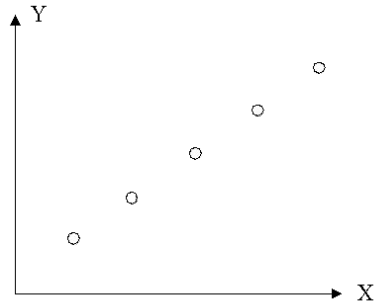
Некоторые переменные связаны между собой линейной зависимостью. Этот закон может быть обусловлен природой анализируемых переменных. В ряде случаев предположение о линейной зависимости допустимо на небольшом интервале изменения аргумента аппроксимирующей функции.

Линейная зависимость двух переменных X и Y может быть установлена при помощи корреляционного анализа. Сила линейной зависимости определяется коэффициентом корреляции r . Если имеется n значений для пар исследуемых переменных, то выборочный коэффициент корреляции определяется следующим образом:

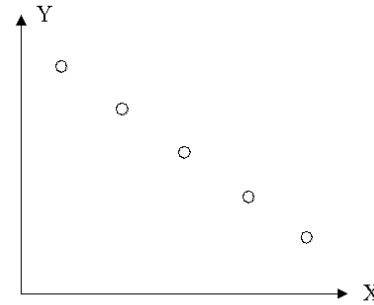
$$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2} \times \sqrt{n \sum Y^2 - (\sum Y)^2}}.$$

Если $r = -1$, то исследуемые величины связаны между собой совершенной отрицательной зависимостью. Напротив, при $r = +1$ между исследуемыми величинами установлена совершенная положительная зависимость. Соответствующие примеры приведены на рисунке – варианты (а) и (б). Два следующих варианта иллюстрируют случаи несовершенной линейной зависимости (положительной и отрицательной). Они наиболее интересны с практической точки зрения. Вариант (е) представляет собой пример нелинейной зависимости. В процессе исследования информации может быть установлено отсутствие зависимости – вариант (ф).

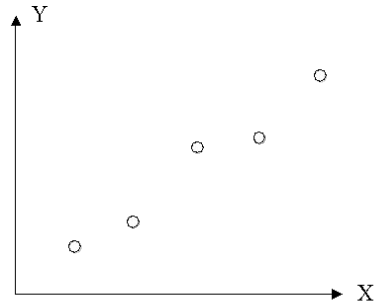
Линейная регрессия (2)



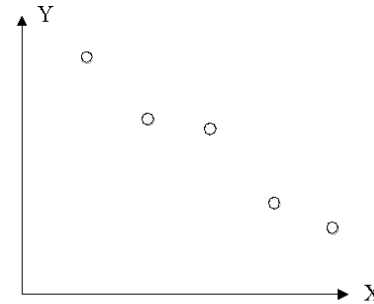
а) совершенная, положительная, линейная



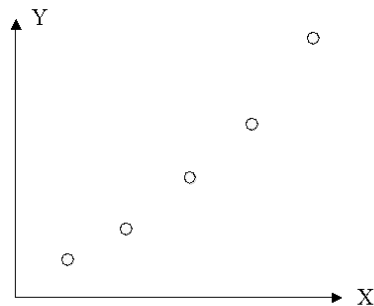
б) совершенная, отрицательная, линейная



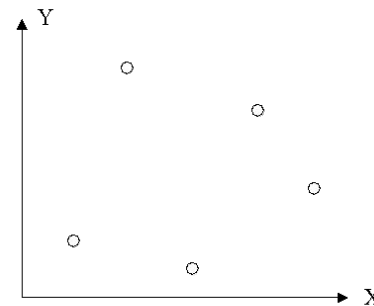
в) несовершенная, положительная, линейная



г) несовершенная, отрицательная, линейная



д) нелинейная



е) зависимость отсутствует

Линейная регрессия (3)

Формула прогностической кривой имеет такой вид:

$$\hat{Y} = b_0 + b_1 X .$$

Величина b_0 называется свободным членом, а b_1 – угловым коэффициентом. Величины b_0 и b_1 определяются методом наименьших квадратов. Соответствующие формулы были приведены в одиннадцатой лекции. Значения b_1 и r пропорциональны друг другу и имеют один и тот же знак.

Для набора пар X и Y в качестве наилучшего приближения выбирается прямая, для которой сумма квадратов отклонений по оси Y минимальна. Эта прямая называется прямой регрессии, а ее уравнение – уравнением регрессии.

После получения уравнения регрессии необходимо определить ошибку аппроксимации набора пар X и Y при помощи кривой. Стандартная ошибка оценки S_{y-x} вычисляется по следующей формуле:

$$S_{y-x} = \sqrt{\frac{\sum(Y - X)^2}{n - 2}} .$$

Метод Бокса-Дженкинса (1)

Изменения объекта или процесса в различные периоды времени часто взаимосвязаны. Эта зависимость измеряется при помощи коэффициента автокорреляции. Для запаздывания на k периодов этот коэффициент r_k между наблюдениями Y_t и Y_{t-k} рассчитывается по такой формуле:

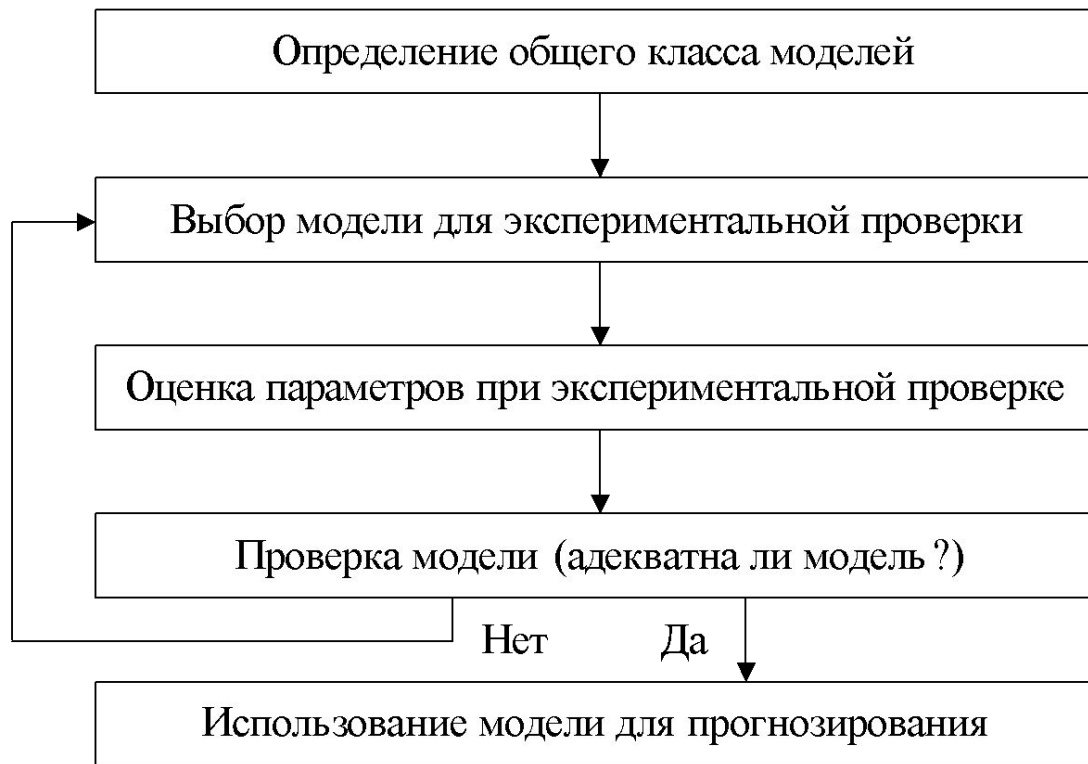
$$r_k = \frac{\sum_{t=k+1}^n (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}.$$

Верхний предел суммирования n определяет общее количество наблюдений.

Для прогнозирования иногда используются модели, в которых анализируется временная структура данных. Модели смешанного авторегрессионного скользящего среднего (AutoRegressive Integrated Moving Average – ARIMA) хорошо описывают как стационарные, так и нестационарные временные ряды. В стационарных рядах результаты изменяются относительно некоего среднего значения. В нестационарных рядах не существует постоянного среднего значения.

В моделях ARIMA для прогнозирования используется информация, которая содержится в исследуемых исходных данных. Модель ARIMA опирается на автокорреляционную структуру данных. В изучение этих моделей значительный вклад внесли G.E.P. Box и G.M. Jenkins.

Метод Бокса-Дженкинса (2)



Метод ARIMA основан на итеративном подходе к выбору приемлемой модели среди возможных вариантов. Такой подход позволяет избежать существенных ошибок, которые возникают вследствие случайного характера отдельных результатов наблюдений.

В настоящее время метод Бокса-Дженкинса реализован практически во всех пакетах программ, используемых для прогнозирования. Появляются модификации данного метода, позволяющие повысить точность прогностических оценок.

Управление процессом (1)

В инфокоммуникационной системе происходят заметные качественные изменения, что отражается на постановке прогностических задач, методах их решения и интерпретации полученных результатов. При постановке прогностических задач следует учитывать два важных аспекта сбора исходных данных:

- в ряде случаев изменяются принципы представления привычных данных, что может привести к появлению ошибок (в частности, такая ситуация иногда возникает при оценке трафика в IP-телефонии);
- для новых видов услуг необходимы данные, которые не собирались из-за отсутствия их практической ценности.

Поэтому управление процессом прогнозирования включает в себя пересмотр принципов сбора и обработки исходных данных. В целом, следует подчеркнуть, что этап сбора и обработки исходных данных очень важен для повышения точности прогнозов, а суть этих операций может часто изменяться.

Управление процессом (2)

Следующий аспект управления процессом прогнозирования заключается в точной постановке соответствующих задач с учетом технических и технологических изменений, свойственных инфокоммуникационной системе. Происходят (иногда – радикальные) изменения в соотношении стоимости отдельных компонентов технических средств. Около тридцати лет назад оперативная память стоила очень дорого. Многие работы были связаны с минимизацией необходимой памяти. Значительная часть прогнозов также была связана с требованиями к объему оперативной памяти. В настоящее время ситуация изменилась. Актуальны стали прогнозы, касающиеся наиболее дорогостоящих компонентов инфокоммуникационной системы.

Прогностические оценки все в большей мере смещаются в область экономических показателей. Это стимулирует использование междисциплинарных подходов к прогнозированию. Кроме того, для некоторых прогнозов актуализируется анализ глобальных трендов, обусловленных формированием информационного общества, интеграцией национальных экономик и им подобными факторами.

Вопросы?