# Практическая работа № 13

# «Построение прогнозов»

**Цель работы:** приобретение навыков прогнозирования и планирования временных рядов.

## Краткие теоретические основания выполнения задания

Регрессионный и корреляционный анализ позволяет установить и оценить зависимость изучаемой случайной величины Y от одной или нескольких других величин X, и делать прогнозы значений Y. Параметр Y, значение которого нужно предсказывать, является зависимой переменной. Параметр X, значения которого нам известны заранее и который влияет на значения Y, называется независимой переменной. Например, X – количество внесенных удобрений, Y – снимаемый урожай; X – величина затрат компании на рекламу своего товара, Y – объем продаж этого товара и т.д.

Корреляционная зависимость Y от X – это функциональная зависимость:

|  |
| --- |
|   (1) |
| где |  *–* среднее арифметическое (условное среднее) всех возможных значений параметра Y, которое соответствует значение X=x. |

Уравнение (1) называется уравнением регрессии Y на X, функция f(x) – регрессией Y на X, а ее график – линией регрессии Y на X.

Основная задача **регрессионного анализа** – установление формы корреляционной связи, то есть вида функции регрессии (линейная, квадратичная, показательная и т.д.).

Метод наименьших квадратов позволяет определить коэффициенты уравнения регрессии таким образом, чтобы точки, построенные по исходным данным (xi, yi), лежали как можно ближе к точкам линии регрессии. Формально это записывается как минимизация суммы квадратов отклонений (ошибок) функции регрессии и исходных точек.

|  |
| --- |
|   (2) |
| где | *–* значение, вычисленное по уравнению регрессии;- отклонение ε (ошибка, остаток) (рис. 1.);n – количество пар исходных данных. |

В регрессионном анализе предполагается, что математическое ожидание случайной величины ε равно нулю и ее дисперсия одинакова для всех наблюдаемых значений Y. Отсюда следует, что рассеяние данных возле линии регрессии должно быть одинаково при всех значениях параметра X.

В случае, показанном на рисунке 2 данные распределяются вдоль линии регрессии неравномерно, поэтому метод наименьших квадратов в этом случае неприменим.

Основная задача корреляционного анализа – оценка тесноты (силы) корреляционной связи. Теснота корреляционной зависимости Y от X оценивается по величине рассеяния значений параметра Y вокруг условного среднего . Большое рассеяние говорит о слабой зависимости Y от X, либо об ее отсутствии и, наоборот, малое рассеяние указывает на наличие достаточно сильной зависимости.



*Рисунок 1 – Понятие отклонения ε для случая линейной регрессии*



*Рисунок 2 – Неравномерное распределение исходных точек вдоль линии регрессии*

Коэффициент детерминации r2 показывает, на сколько процентов (r2\*100%) найденная функция регрессии описывает связь между исходными значениями параметров Y и X:

|  |
| --- |
|   (3) |
| где | *–* объясненная вариация; - общая вариация (рисунок 3). |



*Рисунок 3 – Графическая интерпретация коэффициента детерминации для случая линейной регрессии*

Соответственно, величина (1- r2)\*100% показывает, сколько процентов вариации параметра Y обусловлены факторами, не включенными в регрессионную модель. При высоком (r2≥ 75%) значение коэффициента детерминации можно делать прогноз y\*=f(x\*) для конкретного значения x\*.

Для проведения регрессионного анализа и прогнозирования необходимо:

1. ***построить график*** исходных данных и попытаться зрительно, приближенно определить характер зависимости;
2. ***выбрать вид функции*** регрессии, которая может описывать связь исходных данных;
3. ***определить численные коэффициенты*** функции регрессии;
4. ***оценить силу*** найденной регрессионной зависимости на основе коэффициента детерминации r2;
5. ***сделать прогноз*** (при r2≥ 75%) или сделать вывод о невозможности прогнозирования с помощью найденной регрессионной зависимости. При этом не рекомендуется использовать модель регрессии для тех значений независимого параметра X, которые не принадлежат интервалу, заданному в исходных данных.

**Линейная регрессия.** Коэффициенты линейной регрессии y=a0 +a1x вычисляются по следующим формулам (все суммы берутся по n парам исходных данных):

|  |
| --- |
|   (4) |
|   |

Для удобства вычислений используют вспомогательную таблицу (таблица 1.), в которой рассчитываются необходимые суммы.

*Таблица 1. Вспомогательная таблица для линейной функции*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Заголовки данных |  |  |  |  |  |  |  |
| Промежуточные значения |  |  |  |  |  |  |  |
| … | … | … | … | … | … | … |
|  |  |  |  |  |  |  |
| Сумма ( ) по столбцу |  |  |  |  |  |  |  |

**Пример 1.** Некоторая фирма занимается поставками различных грузов на короткие расстояния внутри города. Перед менеджером стоит задача оценить стоимость таких услуг, зависящую от затраченного на поставку времени. В качестве наиболее важного фактора, влияющего на время поставки, менеджер выбрал пройденное расстояние. Были собраны исходные данные о десяти поставках (таблица 1.15).

*Таблица 2. Исходные данные для примера 1*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Расстояние, миль | 3,5 | 2,4 | 4,9 | 4,2 | 3,0 | 1,3 | 1,0 | 3,0 | 1,5 | 4,1 |
| Время, мин | 16 | 13 | 19 | 18 | 12 | 11 | 8 | 14 | 9 | 16 |

Необходимо построить график исходных данных, определить по нему характер зависимости между расстоянием и затраченным временем, проанализировать применимость метода наименьших квадратов, построить уравнение регрессии, проанализировать силу регрессионной связи и сделать прогноз времени поездки на 2 мили.

***Решение.*** На рисунке 4. построены исходные данные по десяти поездкам.

Помимо расстояния на время поставки влияют пробки на дорогах, время суток, дорожные работы, квалификация водителя, вид транспорта. Построенные точки не находятся точно на линии, что обусловлено описанными выше факторами. Но эти точки собраны вокруг прямой линии, поэтому можно предположить линейную связь между параметрами. Все исходные точки равномерно распределены вдоль предполагаемой прямой линии, что позволяет применить метод наименьших квадратов.

Вычислим суммы, необходимые для расчета коэффициентов линейной регрессии, коэффициента детерминации с помощью таблицы 3.



*Рисунок 4 – График исходных данных для примера 1*

*Таблица 3. Вспомогательная таблица для примера 1*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 3,5 | 16 | 12,25 | 56,00 | 15,223 | 2,634129 | 5,76 |
| 2,4 | 13 | 5,76 | 31,2 | 12,297 | 1,697809 | 0,36 |
| 4,9 | 19 | 24,01 | 93,1 | 18,947 | 28,59041 | 29,16 |
| 4,2 | 18 | 17,64 | 75,60 | 17,085 | 12,14523 | 19,36 |
| 3,0 | 12 | 9,00 | 36,00 | 13,893 | 0,085849 | 2,56 |
| 1,3 | 11 | 1,69 | 14,30 | 9,371 | 17,88444 | 6,76 |
| 1,0 | 8 | 1,00 | 8,00 | 8,573 | 25,27073 | 31,36 |
| 3,0 | 14 | 9,00 | 42,00 | 13,893 | 0,085849 | 0,16 |
| 1,5 | 9 | 2,25 | 13,50 | 9,903 | 13,66781 | 21,16 |
| 4,1 | 16 | 16,81 | 65,60 | 16,819 | 10,36169 | 5,76 |
| ∑=28,9 | ∑=136 | ∑=99,41 | ∑=435,30 | - | 112,4242 | 122,4 |



По формулам (4) вычислим коэффициенты линейной регрессии:

;

.

Таким образом, искомая регрессионная зависимость имеет вид:

  (5)

Наклон линии регрессии а1=2,66 минут на милю – это количество минут, приходящееся на одну милю расстояния. Координата точки пересечения прямой с осью Y а0=5,913 минут – это время, которое не зависит от пройденного расстояния, а обуславливается всеми остальными возможными факторами, явно не учтенными при анализе.

По формуле (28) вычислим коэффициент детерминации:

 или 91,8%.

Таким образом, линейная модель объясняет 91,8% вариации времени доставки. Не объясняется 100% - 91,8% = 8,2% вариации времени поездки, которые обусловлены остальными факторами, влияющими на время поставки, но не включенными в линейную модель регрессии.

Поскольку коэффициент детерминации имеет достаточно высокое значение и расстояние 2 мили, для которого надо сделать прогноз, находится в пределах диапазона исходных данных (см. таблицу 2), то мы можем использовать полученное уравнение линейной регрессии (5) для прогнозирования:

y\* (2 мили) = 5,913+2,660\*2 = 11,2 минут.

При прогнозах на расстояния, не входивших в диапазон исходных данных, нельзя гарантировать справедливость модели (5). Это объясняется тем, что связь между временем и расстоянием может изменяться по мере увеличения расстояния. На время дальних перевозок могут влиять новые факторы такие, как использование скоростных шоссе, остановки на отдых, обед и т.п.

Приблизительным, но самым простым и наглядным способом проверки удовлетворительности регрессионной модели является графическое представление отклонений (рисунок 5).



*Рисунок 5 – График отклонений в примере 1.5*

Отложим отклонений  по оси Y, для каждого значения . Если регрессионная модель близка к реальной зависимости, то отклонения будут носить случайный характер и их сумма будет близка к нулю. В рассмотренном примере .

**Нелинейная регрессия.**

Рассмотрим наиболее простые случаи нелинейной регрессии: гиперболу, экспоненту и параболу. При нахождении коэффициентов гиперболы и экспоненты используют прием приведения нелинейной регрессионной зависимости к линейному виду. Это позволяет использовать для вычисления коэффициентов функции регрессии формулы (4).

*Гипербола*. При нахождении гиперболы  вводят новую переменную , тогда уравнение гиперболы принимает линейный вид . После этого используют формулы (4) для нахождений линейной функции, но вместо значений *xi* используются значения 

; .

При проведении вычислений во вспомогательную таблицу вносятся соответствующие колонки.

*Экспонента*. Для приведения к линейному виду экспоненты  проведем логарифмирование

;

;

.

Введем переменные  и , тогда , откуда следует, что можно применять формулы (29), в которых вместо значений *yi*надо использовать ln *yi*

; .

При этом мы получим численные значения коэффициентов b0 и b1, от которых надо перейти к a0 и a1, используемых в модели экспоненты. Исходя из введенных обозначений и определения логарифма, получаем

, .

*Парабола*. Для нахождения коэффициентов параболы  необходимо решить линейную систему из трех уравнений

,

,



*Оценка силы нелинейной регрессионной связи*. Силы регрессионной связи для гиперболы и параболы определяется непосредственно по формуле (28). При вычислении коэффициента детерминации экспоненты все значения параметра Y (исходные, регрессионные, среднее) необходимо заменить на их логарифмы, например,  - на  и т.д.

**Порядок выполнения задания**

**Пример 1.** По данным о средних доходах на конечное по­требление за десять лет, которые представлены в табл. 1, оцените наличие тренда и в случае положительного ответа постройте трендовую модель.

*Таблица 1. Расходы на конечное потребление, тыс. у.е.*

|  |  |
| --- | --- |
| **Год (t)** | **Расходы *(yt)*** |
| 1-й | 7 |
| 2-й | 8 |
| 3-й | 8 |
| 4-й | 10 |
| 5-й | 11 |
| 6-й | 12 |
| 7-й | 14 |
| 8-й | 16 |
| 9-й | 17 |
| 10-й | 19 |

**Решение**

Для формального определения структуры временного ряда проводится автокорреляционный анализ уровней временного ряда. С этой целью рассчитываются коэффициенты автокорреляции уровней временного ряда. Расчет автокорреляционной функции можно осуществлять на компьютере средствами Excel: **Анализ** данных / **Корреляция.**



На рис. 2 представлены диалоговые окна расчета коэффициентов автокорреляции первого и второго порядков временного ряда *yt.*

|  |  |
| --- | --- |
|  |  |

*Рис. 2. Диалоговые окна расчета коэффициента автокорреляции первого и второго порядков ряда yt*

На рис. 3 представлены результаты расчетов.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ***Столбец1*** | ***Столбец2*** |  |  | ***Столбец1*** | ***Столбец2*** |
| Столбец 1 | 1 |  |  | Столбец 1 | 1 |  |
| Столбец 2 | 0,986854 | 1 |  | Столбец 2 | 0,981221 | 1 |

*Рис. 3. Результаты расчетов коэффициентов автокорреляции.*

Поскольку коэффициенты автокорреляции первых по­рядков (r1, r2) являются высокими, можно предположить на­личие линейного тренда *Т = а + bt.*

Определите уравнение линейного тренда. Для этого:

1. Постройте диаграмму по значениям вре­менного ряда *yt.* (тип *Точечная).*
2. Нажмите правой кнопкой мыши на одной из точек данных на диаграмме. В открывшемся меню необходимо выбрать команду *Добавить линию тренда.* На экране появится диалоговое окно *Линия тренда.*



1. Выберете тип регрессии. Например, *линейная.*
2. Переключитесь на вкладку *Параметры.*  В разделе *Название аппроксимирующей (сглаженной) кривой* установите переключатель *автоматическое*, установите *отображение на диаграмме уравнения* и *величины достоверности аппроксимации.*

Аналогично постройте линию тренда, в качестве функции выбрать степенную.

**Замечание:** Уравнение линейного тренда можно получить, используя инструмент Регрессия Пакета анализа.

**Пример 2.** Провести сглаживание данных задачи 1 и выполнить прогноз на период t=11.

Скопируйте условие предыдущего примера на Лист 2. (Диапазон A1:B11)

С помощью пакета анализа рассчитайте значения скользящего среднего (инструмент **«скользящее среднее»).**

Заполните диалог следующим образом:

* входной интервал - **$В$2:$В$11,**
* интервал - 3,
* выходной интервал -**$С$3,**
* установите флажок **Вывод графика.**

Удалите значения равные **#Н/Д.** Результаты оформите в таблицу с тремя столб­цами: **t, *yt*, Прогноз (скользящ.)**

Скорректируйте построенный график таким образом, чтобы по оси X были значения t (от 1 до 11), по оси У – значения скользящего среднего. График фактических значений *yt*должен быть построен для дней, начиная с 1-го по 10-ый, график прогнозируемых значений должен быть построен для дней начиная с 4-го по 11-ый.

С помощью пакета анализа рассчитайте значения экспоненциального сглаживания (инструмент **«экспоненциальное сглаживание»).**

Заполните диалог следующим образом:

* входной интервал - **$В$2:$В$11,**
* фактор затухания - **0,25,**
* выходной интервал - **$D$2,**
* установите флажок **Вывод графика.**

Удалите значения равные **#Н/Д.**

Продлите значения рассчитанного столбца для получения прогноза на 11-й день. Назовите столбец **Прогноз (экспоненц.).** Скорректируйте построенный график таким образом, чтобы по оси X были значения дней (от 1 до 11), по оси У – спрогнозированные значения. График фактических значений *yt*должен быть построен для дней, начиная с 1-го по 10-ый, график прогнозируемых значений должен быть построен для дней начиная с 2-го по 11-ый.

Сформулируйте экономический смысл полученных моделей. Объясните механизм прогнозирования в каждой их них.

## Задания для самостоятельной работы

**Задание 1.** По данным о выпуске продукции за десять лет, которые представлены в табл. 7, оцените наличие тренда и в случае положительного ответа постройте трендовую модель.

|  |  |
| --- | --- |
| **№ варианта** | **Годы выпуска продукции (t)** |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| 1. | 13,5 | 12,7 | 12 | 11,9 | 11,5 | 11,2 | 10,8 | 10,7 | 10,6 | 10,5 |
| 2. | 251 | 249 | 248 | 246 | 242 | 239 | 235 | 230 | 228 | 225 |
| 3. | 0,91 | 0,87 | 0,85 | 0,82 | 0,79 | 0,75 | 0,7 | 0,66 | 0,62 | 0,6 |

**Задание 2.** По данным задания 1 проведите сглаживание данных и выполните прогноз на период t=11.

## Контрольные вопросы

1. Для чего применяются методы регрессивного и корреляционного анализа?
2. Чем корреляционная зависимость отличается от функциональной?
3. В чем состоит задача регрессионного анализа?
4. В чем состоит задача корреляционного анализа?
5. Как провести регрессионный анализ?
6. Какие виды регрессии существуют?