

Прогнозирование временных рядов: экспоненциальное сглаживание

Лекция 8

Недостатки регрессионного анализа для прогнозирования временных рядов

- ▶ Тренд остается постоянным (прямая, парабола), тогда как в действительности он может измениться.
- ▶ Сезонность одна и та же от начала до конца измерений.

Регрессионные модели плохо себя проявляют при прогнозировании рядов, составляющие которых меняются со временем.

Этих недостатков лишен **метод экспоненциального сглаживания**. кроме того, это очень простой метод.

Метод экспоненциального сглаживания

$$s_0 = x_0$$

$$s_t = \alpha x_t + (1 - \alpha)s_{t-1}, \quad t > 0$$

- ▶ x_t — значение ряда в момент времени t ;
- ▶ s_t — модельное значение в момент времени t ;
- ▶ α — постоянная сглаживания (smoothing factor) $0 < \alpha < 1$.

При расчете прогноза методом экспоненциального сглаживания учитывается отклонение предыдущего прогноза от реального значения ряда

$$s_t = s_{t-1} + \alpha(x_t - s_{t-1})$$

Почему “экспоненциальное” и почему “сглаживание”?

$$\begin{aligned}s_t &= \alpha x_t + (1 - \alpha)s_{t-1} \\&= \alpha x_t + \alpha(1 - \alpha)x_{t-1} + (1 - \alpha)^2 s_{t-2} \\&= \alpha [x_t + (1 - \alpha)x_{t-1} + (1 - \alpha)^2 x_{t-2} + (1 - \alpha)^3 x_{t-3} + \dots + (1 - \alpha)^{t-1} x_1]\end{aligned}$$

Т.е. прогноз строится на основе предыдущих наблюдений, взятых с коэффициентами $1, (1 - \alpha), (1 - \alpha)^2, (1 - \alpha)^3, \dots$, образующими геометрическую прогрессию.

Геометрическая прогрессия является дискретный аналогом **экспоненциальной функции**.

Сглаживание, потому что очередное значение прогноза получается как линейная комбинация предыдущих значений, что позволяет отфильтровать (сгладить) сезонную составляющую.

Поскольку $0 < \alpha < 1$, то вес наблюдений в прогнозе уменьшается с их “возрастом” — более старые наблюдения имеют меньший вес.

Двойное экспоненциальное сглаживание

Если ряд имеет тренд, значения такого ряда записываются как

$$x_t = a + bt$$

Тогда

$$b = x_t - x_{t-1}$$

Позволим теперь тренду меняться: $b \rightarrow b_t$.

Для $t = 0, 1$

$$s_1 = x_1$$

$$b_1 = x_1 - x_0$$

Для $t > 2$

$$s_t = \alpha x_t + (1 - \alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1 - \beta)b_{t-1}$$

- ▶ α — показатель сглаживания для данных, $0 < \alpha < 1$;
- ▶ β — показатель сглаживания для тренда, $0 < \beta < 1$.

$$b_t = b_{t-1} + \beta(s_t - (s_{t-1} + b_{t-1}))$$

Локальный тренд

b_t называют **локальным трендом**. Он позволяет ряду подстраиваться под изменение тренда. Но если тренд будет меняться слишком часто, то такой метод прогноза будет работать плохо.

Может теперь добавить в ряд еще и сезонность?

Тройное экспоненциальное сглаживание или метод Холта-Уинтерса (Holt-Winters)

С мультипликативной сезонной составляющей:

$$s_0 = x_0$$

$$s_t = \alpha \frac{x_t}{c_{t-L}} + (1 - \alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1 - \beta)b_{t-1}$$

$$c_t = \gamma \frac{x_t}{s_t} + (1 - \gamma)c_{t-L}$$

- ▶ c_t — локальная сезонность;
- ▶ γ — показатель сглаживания для сезонных изменений, $0 < \gamma < 1$;
- ▶ L — длина периода (сезона).

Метод Холта-Уинтерса С аддитивной составляющей

$$s_0 = x_0$$

$$s_t = \alpha(x_t - c_{t-L}) + (1 - \alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1 - \beta)b_{t-1}$$

$$c_t = \gamma(x_t - s_{t-1} - b_{t-1}) + (1 - \gamma)c_{t-L}$$

Итак

Чтобы прогнозировать значения ряда с трендом и сезонной составляющей нам нужно знать значения трех параметров: α , β , γ .

Как найти α , β и γ

Ищем величины α , β , γ и начальных значений параметров сглаживания, которые минимизируют выражение

$$SSE = \sum_{t=1}^T (x_t - s_t)^2 = \sum_{t=1}^T e_t^2$$

$e_t = x_t - s_t$ — ошибки прогноза.

$t = 1, \dots, T$.

Область применения

Краткосрочные прогнозы — на несколько месяцев вперед.

Пример: прогнозирование международных авиаперевозок

```
# Шаг 0. Прочитаем данные
ser.g.01 <- read.table("week_10/data/series_g.csv",
                      header=T, sep=";")

# Проверка: импортировали правильно?
summary(ser.g.01)

# Шаг 1. Создадим временной ряд
ser.g.02 <- ts(data = ser.g.01[,2], frequency = 12,
               start = c(1949, 1) )

# Проверка: что получилось?
ser.g.02
```

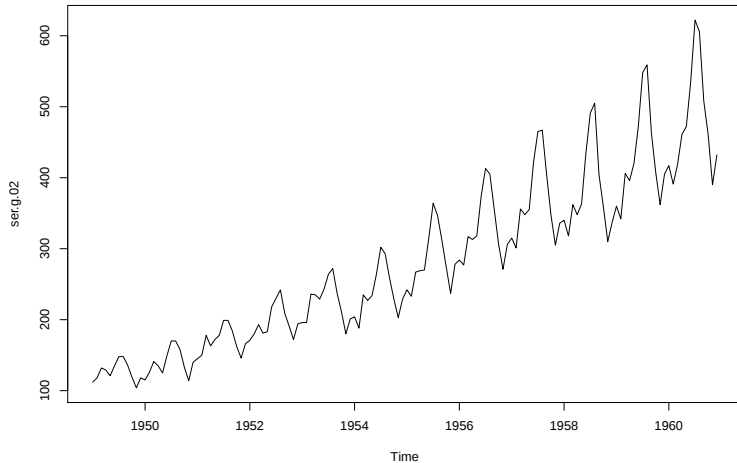
ts() создает временной ряд, т.е. вектор с информацией о периоде сезонной составляющей (frequency = 12) и дате старта в формате c(год, месяц).

Временной ряд, созданный ts()

##		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
##	1949	112	118	132	129	121	135	148	148	136	119	104	118
##	1950	115	126	141	135	125	149	170	170	158	133	114	140
##	1951	145	150	178	163	172	178	199	199	184	162	146	166
##	1952	171	180	193	181	183	218	230	242	209	191	172	194
##	1953	196	196	236	235	229	243	264	272	237	211	180	201
##	1954	204	188	235	227	234	264	302	293	259	229	203	229
##	1955	242	233	267	269	270	315	364	347	312	274	237	278
##	1956	284	277	317	313	318	374	413	405	355	306	271	306
##	1957	315	301	356	348	355	422	465	467	404	347	305	336
##	1958	340	318	362	348	363	435	491	505	404	359	310	337
##	1959	360	342	406	396	420	472	548	559	463	407	362	405
##	1960	417	391	419	461	472	535	622	606	508	461	390	432

Рассмотрим график

```
plot(ser.g.02)
```



Проведем экспоненциальное сглаживание

```
ser.g.HW <- HoltWinters(ser.g.02, seasonal = "mult")
```

Поиск параметров производится функцией оптимизации `optim()` (метод Нелдера-Мида).

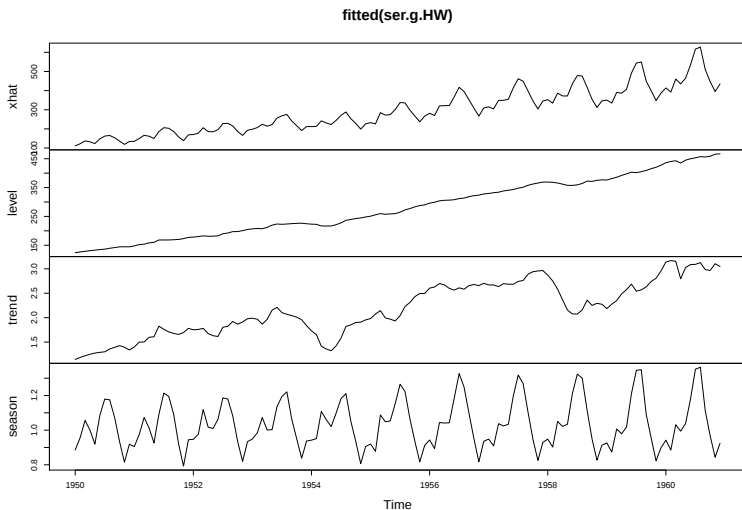
- ▶ `optim.start` — начальные значения параметров: `c(alpha = 0.3, beta = 0.1, gamma = 0.1)`
- ▶ `optim.control` — передача списка параметров в функцию `optim`.

Простое экспоненциальное сглаживание:

```
HoltWinters(ser.g.02, alpha=0.1, beta=FALSE, gamma=FALSE)
```

Визуализация составляющих ряда

```
plot(fitted(ser.g.HW))
```



Прогнозируем на год вперед

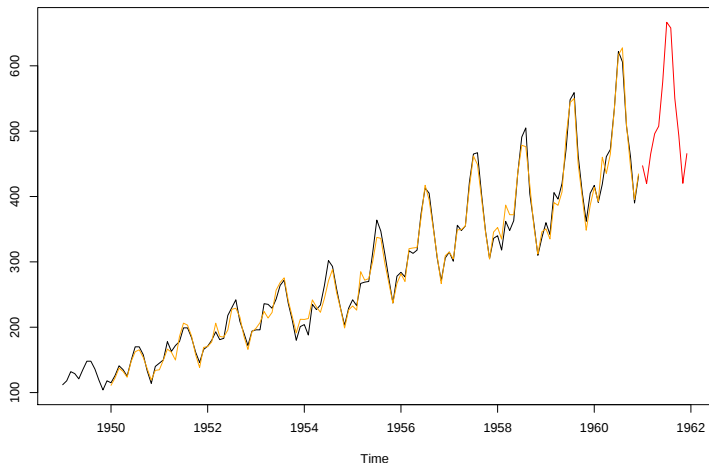
`n.ahead` — на сколько месяцев вперед прогнозировать

```
ser.g.predict <- predict(ser.g.HW, n.ahead=12)  
ser.g.predict
```

##	Jan	Feb	Mar	Apr	May	Jun	Jul
## 1961	447.0559	419.7123	464.8671	496.0839	507.5326	575.4509	666.5923
##	Aug	Sep	Oct	Nov	Dec		
## 1961	657.9137	550.3088	492.9853	420.2073	465.6345		

Строим график исходного ряда, подгонки и прогноза

```
ts.plot(ser.g.02, ser.g.predict, ser.g.HW$fitted[,1],  
        col=c("black", "red", "orange"))
```



Пример. Продажи красных вина в Австралии, прогноз на 8 месяцев

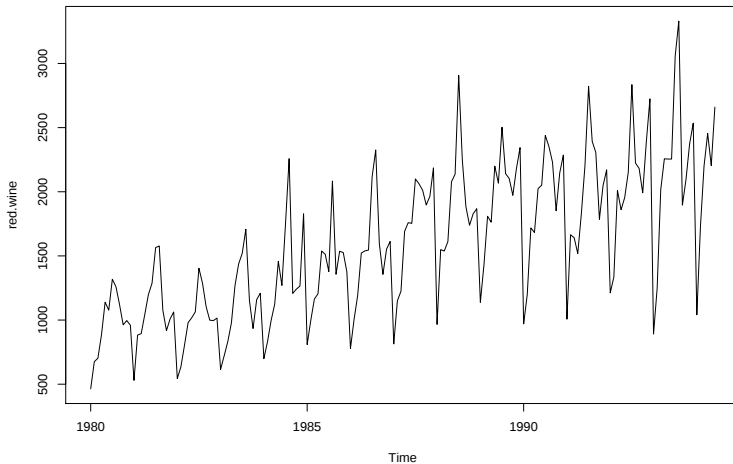
```
# Шаг 0. Прочитаем данные  
# Посмотрим, как выглядит файл  
file.show("wine_Austral.dat")  
wine.01 <- read.table("wine_Austral.dat")  
# Проверка: импортировали правильно?  
summary(wine.01)
```

Шаг 1. Создадим временной ряд

```
red.wine <- ts(data = wine.01[ ,4],  
               frequency = 12, start = c(1980, 1) )  
# Проверка: что получилось?  
red.wine
```

Шаг 2. Рассмотрим график

```
plot(red.wine)
```

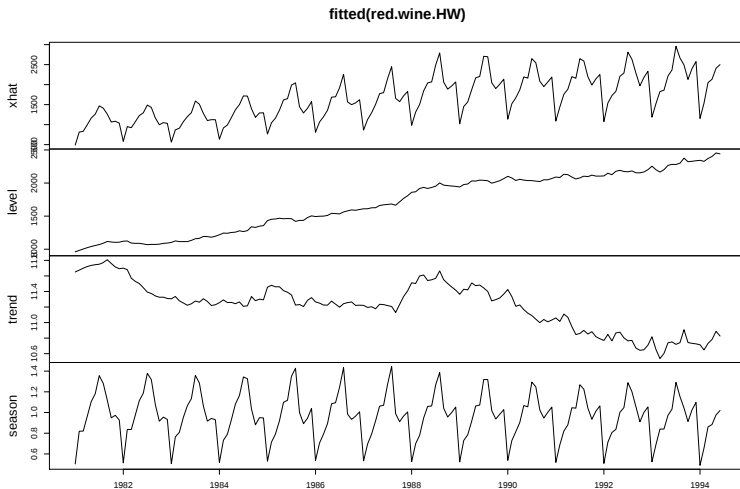


Ключевые вопросы

1. Есть ли у ряда тренд?
2. Есть ли сезонность?
3. Меняет ли ряд свой характер?
4. Есть ли в данных выбросы?

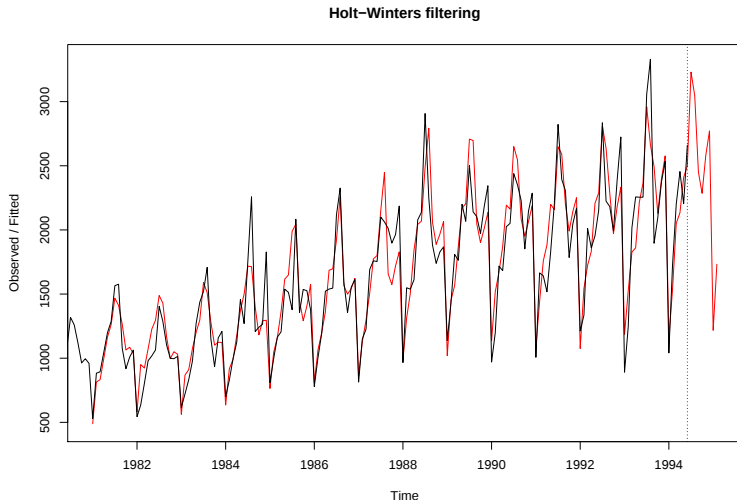
Шаги 3-4. Проведем экспоненциальное сглаживание и посмотрим на результаты

```
red.wine.HW <- HoltWinters(red.wine, seasonal = "mult")  
plot(fitted(red.wine.HW))
```



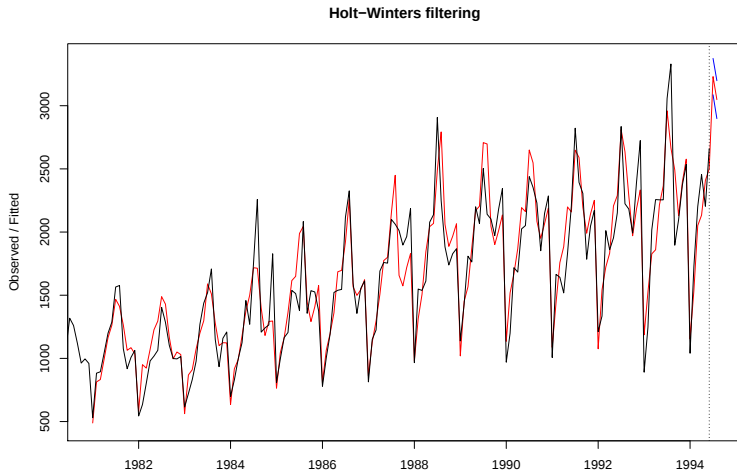
Шаг 5. Прогнозируем на 8 мес.

```
red.wine.predict <- predict(red.wine.HW, n.ahead=8)  
plot(red.wine.HW, red.wine.predict)
```



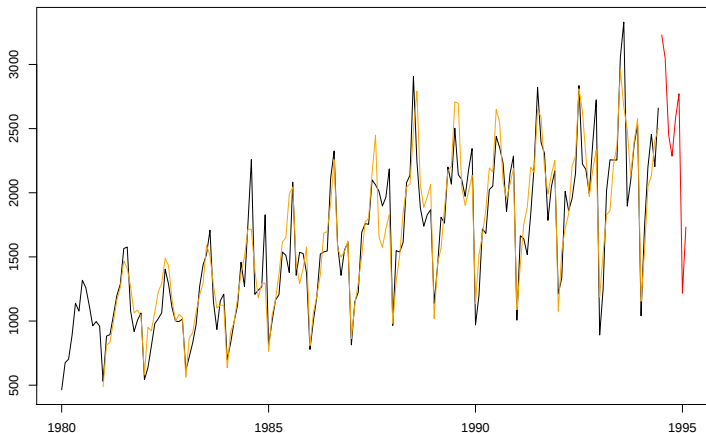
Шаг 5а. Посмотрим доверительный интервал прогноза

```
forecast <- predict(red.wine.HW, n.ahead=2,  
                    prediction.interval = T, level = 0.95)  
plot(red.wine.HW, forecast)
```



Шаг 6. Строим график исходного ряда, подгонки и прогноза

```
ts.plot(red.wine, red.wine.predict, red.wine.HW$fitted[,1],  
        col=c("black", "red", "orange"))
```



Когда экспоненциальное сглаживание проигрывает линейной регрессии

У коротких временных рядов (несколько сезонов) для качественного экспоненциального сглаживания недостаточно данных, тогда как предположение о постоянстве тренда и сезонных поправок вполне работоспособно.

Как оценить ошибку прогноза?

Алгоритм состоит в следующем:

1. Перед проведением анализа исходный временной ряд укорачивается на 7-10% (последний месяц, квартал и т.п.), и этот “хвост” ряда сохраняется для дальнейшего анализа.
2. Строится прогноз “укороченного” ряда и результат сравнивается с отложенными данными.

Если под конец наблюдаемого периода на поведение ряда не повлияли новые факторы, то мера ошибки, полученной при сопоставлении, будет оценкой ошибки на будущий (прогнозный) период.

Т.е., если прогноз на 2—5% отличается от наблюдаемых данных на “хвосте” ряда, то можно ожидать, что эта ошибка сохранится и в прогнозный период.

Как выбрать метод прогноза?

Аналогично, то есть:

- ▶ отрезать “хвост” ряда и проверить на нем прогнозы, сделанные разными методами. Какой метод даст при этом меньшую сумму квадратов ошибок (MSE) или сумму модулей ошибок (на отделенном заранее “хвосте”), тот и лучше.

Эта процедура называется **перекрестная проверка** или **кросс-валидация** (cross validation, cross-validation).

Дополнительные материалы

- ▶ **Пакет forecast** — наиболее популярный пакет R для прогнозирования одномерных временных рядов. От Роба Хиндмана (Rob Hyndman) и компании.
- ▶ **Quick-R. Time Series and Forecasting** — сводка функций R, посвященных прогнозированию временных рядов, с сайта Р. Кабакова.