



Институт технологий управления

Программные средства анализа данных

01.03.05 Статистика

Профиль «Анализ данных в бизнесе и экономике»

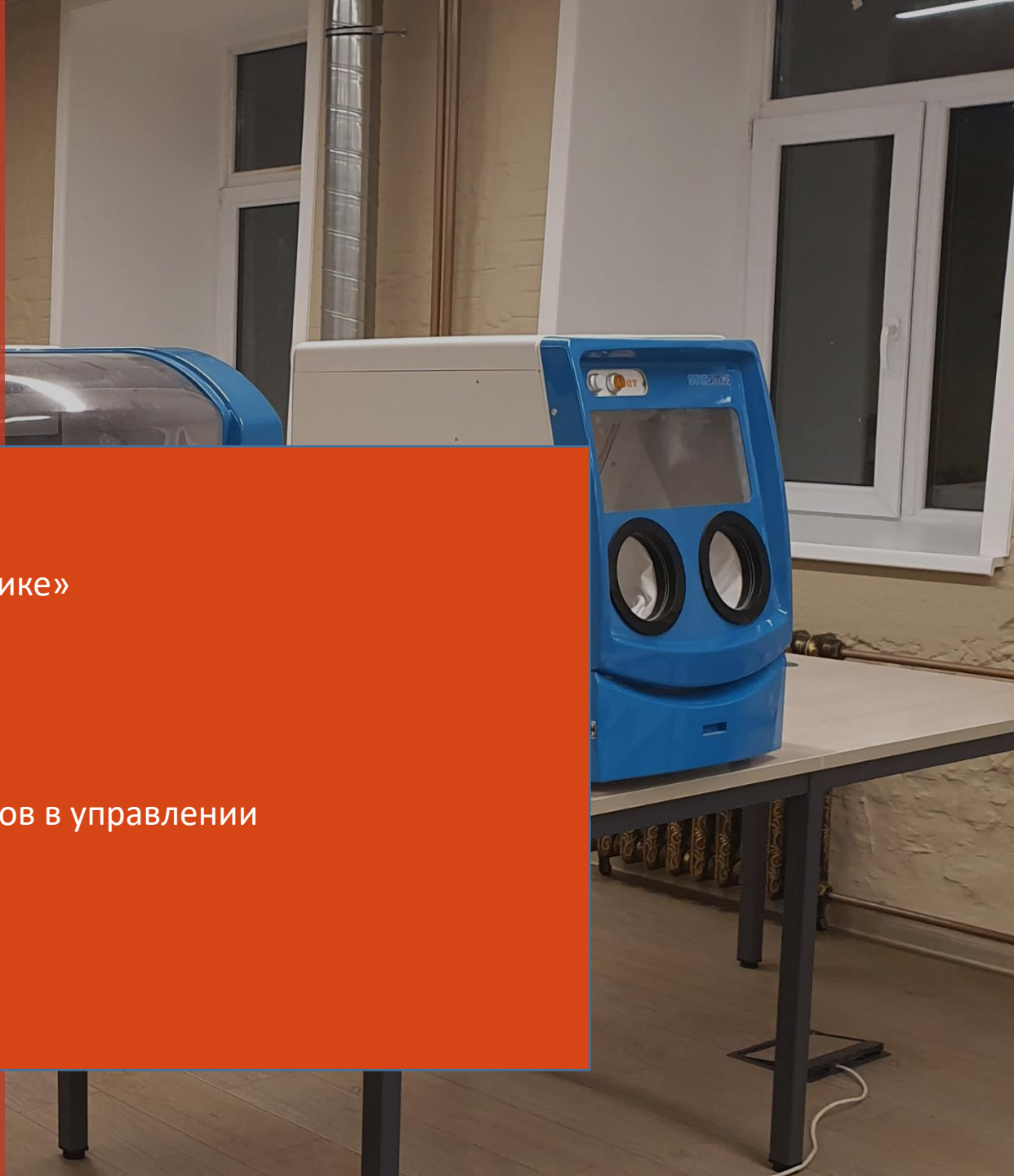
Квалификация «бакалавр»

Бурцева Татьяна Александровна

Профессор кафедры статистики и математических методов в управлении

Burceva_t@mirea.ru

Москва, 2022



Тема 7. Регрессия во временных рядах. Регрессия по панельным данным.

План лекции

1. Модели тренда.
2. Модели сезонности.
3. Модели тренда и сезонности.
4. Модели с эффектами.

1. Модели тренда

В краткосрочном периоде прогнозирования на экономический объект действует немало случайных факторов, ослабляющих определяющие тенденции его развития. Поэтому использование экстраполяции в прогнозировании имеет в своей основе предположение о том, что рассматриваемый процесс изменения той или иной экономической переменной представляет собой сочетание двух составляющих — x_t **регулярной** (детерминированной неслучайной) и ε_t **случайной**. Тогда временной ряд экономического показателя y_t может быть представлен в следующем виде: $y_t = x_t + \varepsilon_t$.

Регулярная составляющая называется **тенденцией, трендом**. Регулярная составляющая (тренд) x_t характеризует существующую динамику развития процесса в целом, случайная составляющая ε_t отражает случайные колебания или **шумы процесса**. Обе составляющие определяются каким-либо функциональным механизмом, характеризующим их поведение во времени.

Задача прогноза состоит в определении вида экстраполирующей функции x_t и ε_t (на основе исходных эмпирических данных) и параметров выбранной функции – модели тренда.

В R существует большое количество пакетов для анализа временных рядов (forecast, prophet)

Моделирование экономической динамики



Прогнозные расчёты на основе трендовых моделей строятся в два этапа. На первом (формальном) — выявляют при помощи статистических методов закономерности прошлого развития и переносят (экстраполируют) их на некоторый период будущего. На втором — производится корректировка полученного прогноза с учётом результатов содержательного анализа текущего состояния и действия экономического механизма на период прогнозирования.

При моделировании экономической динамики осуществляют следующие основные операции:

- 1) сглаживание (выравнивание) исходного ряда в целях более чёткого выявления тенденции развития исследуемого процесса (методом скользящего среднего, экспоненциального сглаживания и др.);
- 2) определение наличия тренда, т.е. изменения, определяющего общее направление развития, основную тенденцию временного ряда (на основе графического метода, использования аппарата теории вероятностей и математической статистики);
- 3) отбор одной или нескольких кривых роста;
- 4) определение их параметров;
- 5) оценку адекватности и точности трендовых моделей по критериям серий, пиков, на основе исследования показателей асимметрии и эксцесса, t-критерия Стьюдента, d-критерия Дарбина-Уотсона, среднего квадратического отклонения, относительной ошибки аппроксимации, коэффициента сходимости и др.

Виды моделей тренда



Модель тренда может иметь различный вид. Её выбор в каждом конкретном случае осуществляется по целому ряду статистических критериев. Наибольшее распространение при построении трендовых моделей экономических процессов получили **полиномиальные, экспоненциальные и S-образные** кривые роста.

Простейшие полиномиальные кривые роста имеют вид:

$y_t = b + m_1 t$ (полином первой степени);

$y_t = b + m_1 t + m_2 t^2$ (полином второй степени);

$y_t = b + m_1 t + m_2 t^2 + m_3 t^3$ (полином третьей степени) и т.д.

Такие кривые роста можно использовать для аппроксимации (приближения) и прогнозирования экономических процессов, в которых последующее развитие не зависит от достигнутого уровня.

Использование **экспоненциальных кривых роста** предполагает, что дальнейшее развитие зависит от достигнутого уровня (например, прирост зависит от значения функции).

В экономических расчётах чаще всего применяют две **разновидности экспоненциальных (показательных) кривых**:

1) простая экспонента вида $y_t = b m^t$;

2) модифицированная экспонента вида $y_t = a + b m^t$.

В экономике достаточно часто встречаются процессы, которые сначала растут медленно, затем ускоряются, а затем снова замедляют свой рост, стремясь к какому-либо пределу. Такой тип развития характерен, например, для спроса на некоторые новые товары. Для моделирования подобных процессов используют **S-образные кривые роста**, среди которых выделяют кривую Гомперца ($y_t = k a^{b^t}$), кривую Перла – Риды ($1/y_t = k + a b^t$) и логистическую кривую ($y_t = k/(1 + b e^t)$).

Построение трендов в R по численности населения России (подготовка переменных)

#Ввод из файла din.xlsx

```
library(readxl)
```

```
types = c("text", rep("numeric", 1))
```

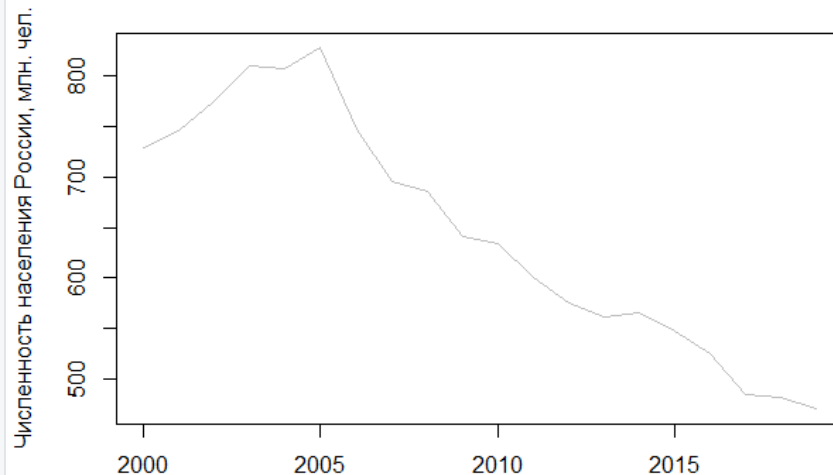
```
s <- as.data.frame(read_excel("C:/Users/компьютер/Documents/din.xlsx", 1,  
                             col_types = types))
```

#добавление номера периода как переменной в data frame s

```
for (i in 1:length(s$год)){s$t[i]=i}
```

#построение графика

```
plot(s$год, s$s, col="grey", type="l", xlab="",  
     ylab="Численность населения России, млн. чел.")
```



	год	s	t
1	2000	729.1	1
2	2001	745.4	2
3	2002	775.6	3
4	2003	810.8	4
5	2004	807.9	5
6	2005	827.8	6
7	2006	747.2	7
8	2007	695.8	8
9	2008	685.4	9
10	2009	640.7	10
11	2010	634.0	11
12	2011	600.9	12
13	2012	575.7	13
14	2013	560.9	14
15	2014	565.6	15
16	2015	546.7	16
17	2016	525.3	17
18	2017	484.5	18
19	2018	482.2	19
20	2019	470.0	20

	год	s
1	2000	729.1
2	2001	745.4
3	2002	775.6
4	2003	810.8
5	2004	807.9
6	2005	827.8
7	2006	747.2
8	2007	695.8
9	2008	685.4
10	2009	640.7
11	2010	634.0
12	2011	600.9
13	2012	575.7
14	2013	560.9
15	2014	565.6
16	2015	546.7
17	2016	525.3
18	2017	484.5
19	2018	482.2
20	2019	470.0

А	В
2005	827,8
2006	747,2
2007	695,8
2008	685,4
2009	640,7
2010	634,0
2011	600,9
2012	575,7
2013	560,9
2014	565,6
2015	546,7
2016	525,3
2017	484,5
2018	482,2
2019	470,0

Построение полиномиальных трендов на графике

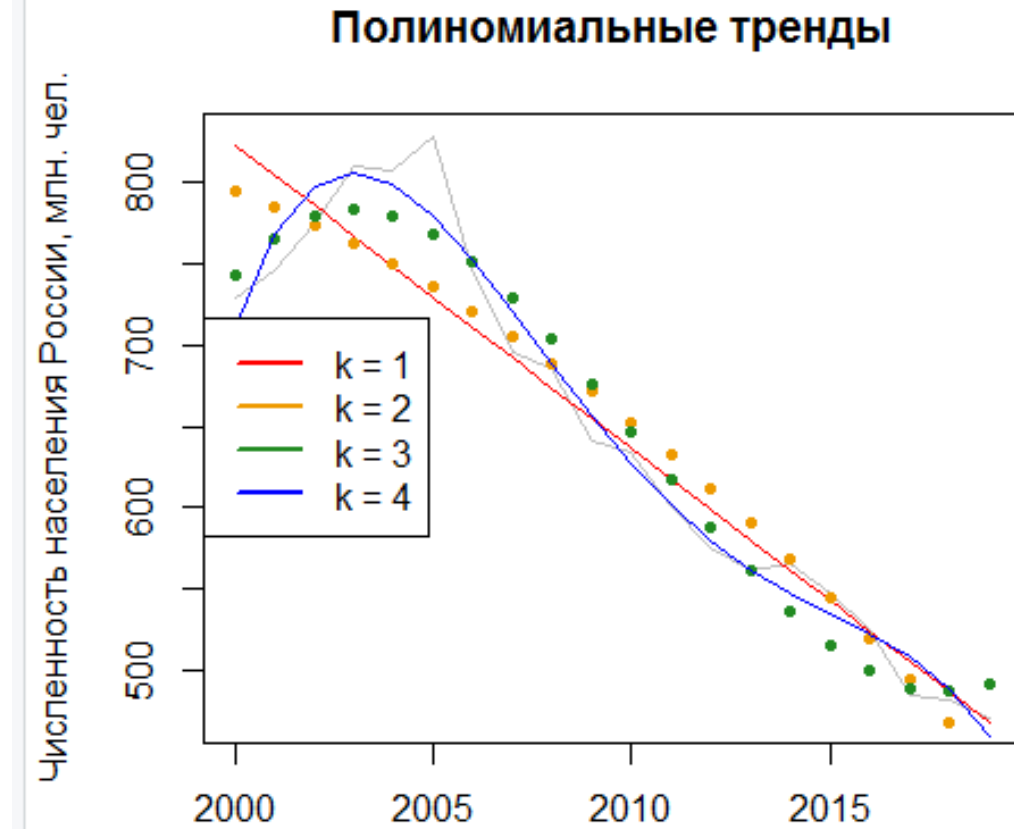
```
points(s[,1], predict(lm(s$s~poly(s$t, k=1,  
raw = TRUE))), pch=20, col="red", type="l")
```

```
points(s[,1], predict(lm(s$s~poly(s$t, k=2,  
raw = TRUE))), pch=20, col="orange2",  
type="p")
```

```
points(s[,1], predict(lm(s$s~poly(s$t, k=3,  
raw = TRUE))), pch=20, col="forestgreen",  
type="p")
```

```
points(s[,1], predict(lm(s$s~poly(s$t, k=4,  
raw = TRUE))), pch=20, col="blue", type="l")
```

```
legend("left", c("k = 1", "k = 2", "k = 3", "k =  
4"), col = c("red", "orange2",  
"forestgreen", "blue" ), lwd = 2)
```



Смотрим уравнение и надёжность трендов

#Для первого тренда (линейный)

```
ss = lm(s$s ~ poly(s$t, k=1, raw = TRUE))
```

```
summary(ss)
```

```
call:
lm(formula = s$s ~ poly(s$t, k = 1, raw = TRUE))

Residuals:
    Min       1Q   Median       3Q      Max
-94.140 -17.272  -0.469   6.080  98.068

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      841.942      19.235   43.77  < 2e-16 ***
poly(s$t, k = 1, raw = TRUE) -18.702       1.606  -11.65 8.15e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.41 on 18 degrees of freedom
Multiple R-squared:  0.8828,    Adjusted R-squared:  0.8763
F-statistic: 135.6 on 1 and 18 DF,  p-value: 8.151e-10
```


Смотрим уравнение и надёжность трендов

```
Call:
lm(formula = s$s ~ poly(s$t, k = 2, raw = TRUE))
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-65.936 -29.846  -2.908  17.609  91.636
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    803.8420    29.4098   27.332 1.73e-15 ***
poly(s$t, k = 2, raw = TRUE)1  -8.3108     6.4500   -1.288   0.215
poly(s$t, k = 2, raw = TRUE)2  -0.4948     0.2983   -1.658   0.116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 39.53 on 17 degrees of freedom
Multiple R-squared:  0.8992,    Adjusted R-squared:  0.8873
F-statistic: 75.79 on 2 and 17 DF,  p-value: 3.394e-09
```

```
lm(formula = s$s ~ poly(s$t, k = 3, raw = TRUE))
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-35.894 -17.122  -4.609  26.501  59.757
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    708.31373    30.28515   23.388 8.46e-14 ***
poly(s$t, k = 3, raw = TRUE)1  40.43327    12.18562    3.318 0.004350 **
poly(s$t, k = 3, raw = TRUE)2  -6.15853     1.33125   -4.626 0.000280 ***
poly(s$t, k = 3, raw = TRUE)3   0.17980     0.04174    4.308 0.000542 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 27.73 on 16 degrees of freedom
Multiple R-squared:  0.9533,    Adjusted R-squared:  0.9446
F-statistic: 108.9 on 3 and 16 DF,  p-value: 7.375e-11
```

```
Call:
lm(formula = s$s ~ poly(s$t, k = 4, raw = TRUE))
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-25.517  -8.580  -0.880   9.432  48.318
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    619.850181    31.040988   19.969 3.23e-12 ***
poly(s$t, k = 4, raw = TRUE)1  109.636474    19.402032    5.651 4.61e-05 ***
poly(s$t, k = 4, raw = TRUE)2  -20.162208     3.631696   -5.552 5.54e-05 ***
poly(s$t, k = 4, raw = TRUE)3   1.199638     0.256897    4.670 0.000302 ***
poly(s$t, k = 4, raw = TRUE)4  -0.024282     0.006075   -3.997 0.001166 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 19.93 on 15 degrees of freedom
Multiple R-squared:  0.9774,    Adjusted R-squared:  0.9714
F-statistic: 162.1 on 4 and 15 DF,  p-value: 3.779e-12
```

Полином четвертой
степени лучший

Пример 2. Тренд занятых в Калужской области

```
#загрузка годовых данных по численности занятого населения РФ с 2014 по 2019 гг по регионам РФ
```

```
library(readxl)
```

```
types = c("text", rep("numeric", 6))
```

```
yt <- as.data.frame(read_excel("C:/Users/компьютер/Documents/employed.xlsx", 1,  
                             col_types = types))
```

```
#данные по Калужской области о численности занятого населения
```

```
yt1=yt[6,2:7]
```

```
#как достать уровни ряда
```

```
y=c(yt1[1,1],yt1[1,2],yt1[1,3], yt1[1,4], yt1[1,5], yt1[1,6])
```

```
t=c(1:length(yt1)) # присваиваем номера уровней ряда
```

```
fm=lm(y ~ t) # вычисляем линейный тренд
```

```
summary(fm) # смотрим модель
```

```
plot(fm) # строим график
```

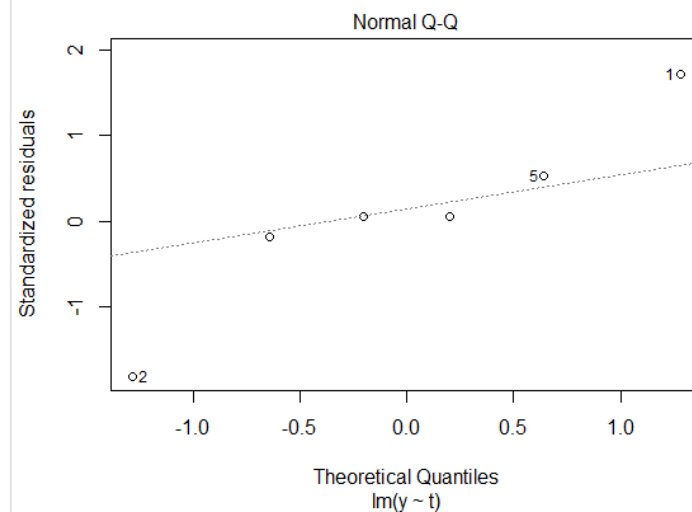
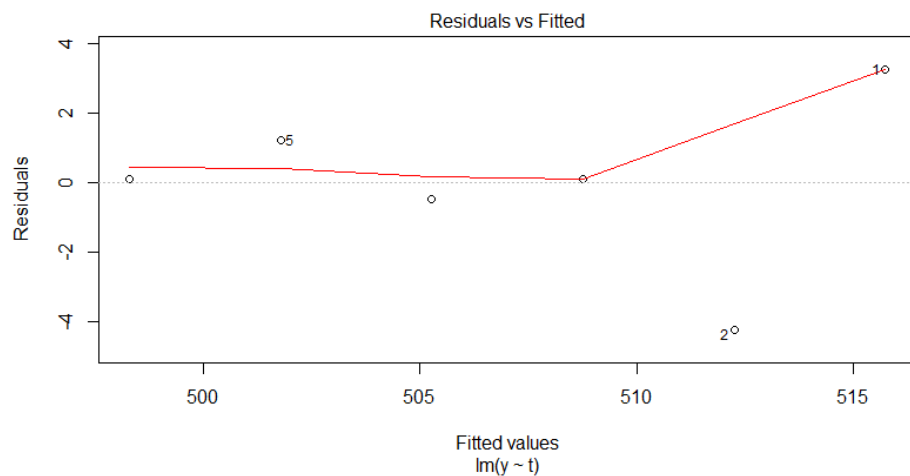
```
Call:
lm(formula = y ~ t)

Residuals:
    1     2     3     4     5     6 
3.2501 -4.2153  0.1132 -0.4732  1.2173  0.1079 

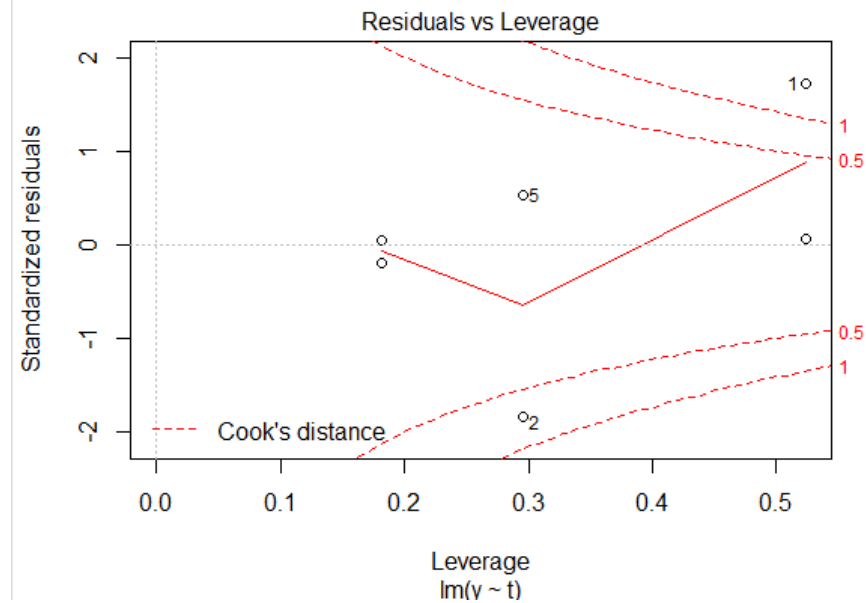
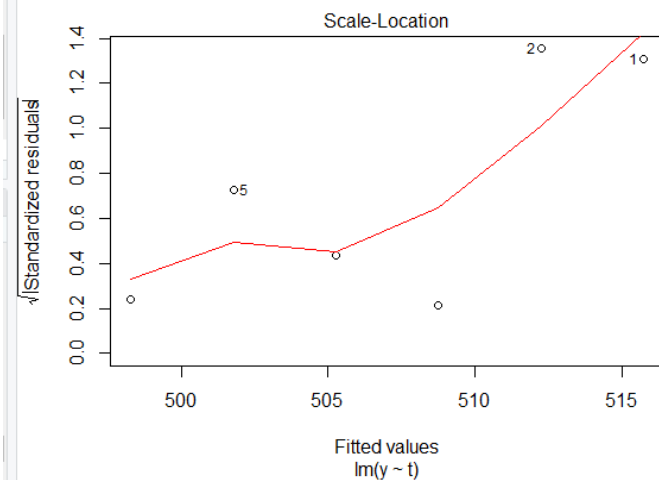
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  519.2354     2.5522  203.449  3.5e-09 ***
t            -3.4905     0.6553   -5.326  0.00598 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.741 on 4 degrees of freedom
Multiple R-squared:  0.8764,    Adjusted R-squared:  0.8455 
F-statistic: 28.37 on 1 and 4 DF, p-value: 0.00598
```

Диагностические диаграммы остатков



- `Plot(fm)`

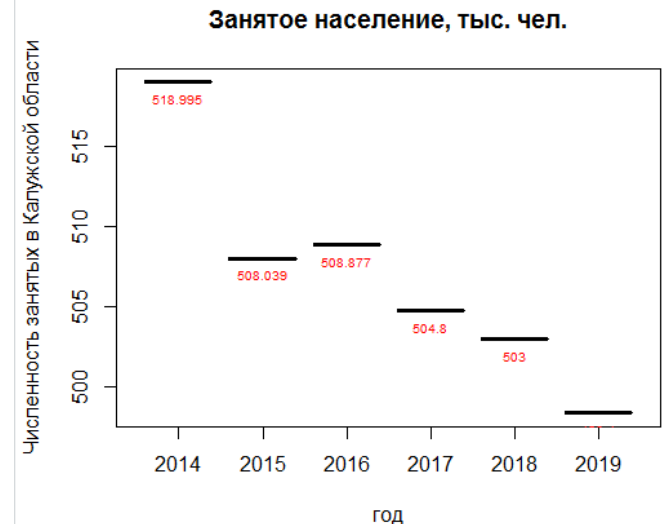


Пример построения тренда численности занятых в Калужской области

```
55 #загрузка годовых данных по численности занятого населения РФ с 2014 по 2019 гг по регионам
56 library(readxl)
57 types = c("text", rep("numeric", 6))
58 yt <- as.data.frame(read_excel("C:/Users/компьютер/Documents/employed.xlsx", 1,
59                               col_types = types))
60 #данные по калужской области о численности занятого населения
61 yt1=yt[6,2:7]
62 #как достать уровни ряда
63 y=c(yt1[1,1],yt1[1,2],yt1[1,3], yt1[1,4], yt1[1,5], yt1[1,6])
64 t=c("2014", "2015", "2016", "2017", "2018", "2019")
65 i=c(1,2,3,4,5,6)
66 #число столбцов
67 ncol(yt1)
68 #число строк
69 nrow(yt1)
70 df=data.frame(t,y,i)
71 plot( df$t, df$y, xlab="год", ylab="численность занятых в калужской области", main = "Занят
72
73 text(df$t, df$y,
74      df$y,
75      cex=0.6, pos=1, col="red")
```

регионы	2014	2015	2016	2017	2018	2019
1 Белгородская область	764.455	754.042	756.835	757.9	752.6	754.1
2 Брянская область	557.766	547.669	540.643	530.2	523.0	508.6
3 Владимирская область	669.711	664.413	647.434	640.6	628.2	635.8
4 Воронежская область	1117.774	1092.535	1094.752	1102.1	1110.2	1106.4
5 Ивановская область	455.875	451.493	447.059	456.3	444.9	443.3
6 Калужская область	518.995	508.039	508.877	504.8	503.0	498.4
7 Костромская область	307.561	299.406	293.153	290.8	282.2	276.8
8 Курская область	529.530	520.324	520.554	519.6	510.8	505.5
9 Липецкая область	599.317	565.151	565.450	565.8	566.1	565.1
10 Московская область	3405.262	3366.883	3376.991	3450.2	3385.7	3437.1
11 Орловская область	340.605	335.905	330.187	321.1	314.5	298.7
12 Рязанская область	513.557	504.808	505.504	511.0	498.3	494.6
13 Смоленская область	470.036	460.832	443.850	445.9	432.5	411.4
14 Тамбовская область	502.179	499.777	492.131	482.4	466.0	454.1

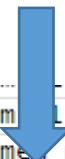
	t	y	i
1	2014	518.995	1
2	2015	508.039	2
3	2016	508.877	3
4	2017	504.800	4
5	2018	503.000	5
6	2019	498.400	6



Команда **ts** автоматически преобразует переменные в динамический ряд

- Создаём динамический ряд

`y3=ts(y, start=c(2014,1), end=c(2019,1), frequency = 1)`

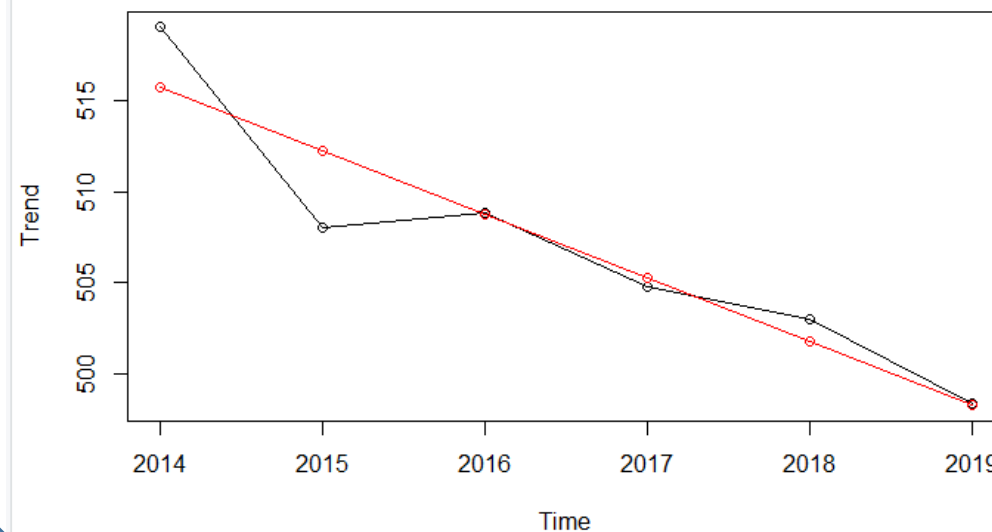



..
y	num [1:6] 519 508 509 505 503 ...
y2	Name num [1:6] 516 512 509 505 502 ...
y3	Time-Series [1:6] from 2014 to 2019: 519 508 509 505 503 ...

```
types = c("text", rep("numeric", 6))
yt <- as.data.frame(read_excel("C:/Users/компьютер/Documents/employed.xlsx", 1,
                              col_types = types))

#данные по калужской области о численности занятого населения
yt1=yt[6,2:7]

#как достать уровни ряда
y=c(yt1[1,1],yt1[1,2],yt1[1,3], yt1[1,4], yt1[1,5], yt1[1,6])
t=c(1:length(yt1))
y3=ts(y, start=c(2014,1), end=c(2019,1), frequency = 1)
fm=lm(y3 ~ t)
summary(fm)
y2=predict(fm)
library(graphics)
library(stats)
sp=c("Занятые", "линейный тренд")
ts.plot(y3,y2, ylab="Trend", col=1:2, type="o")
```

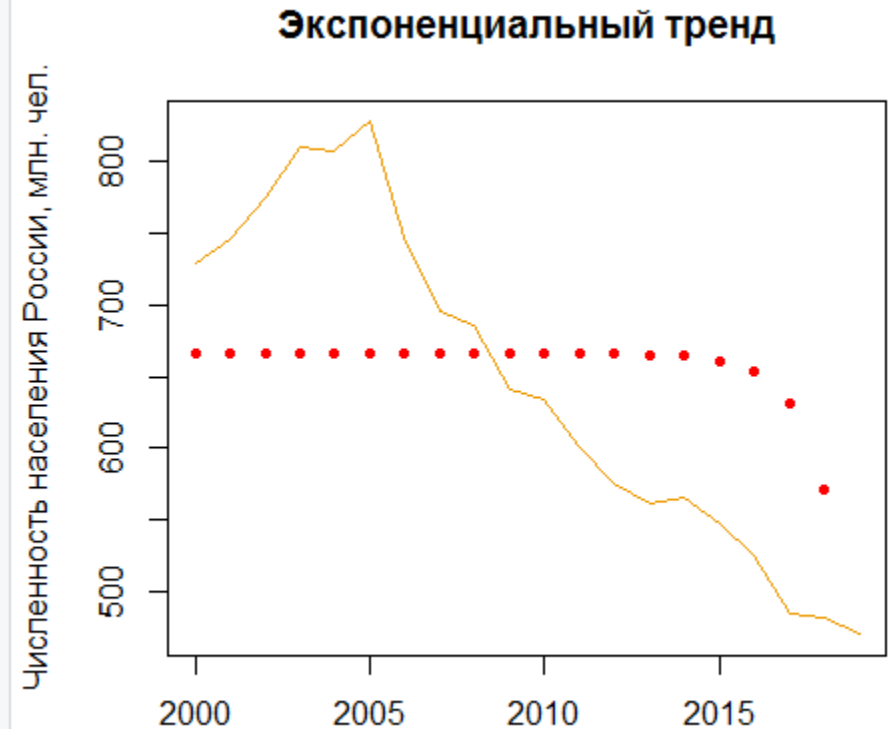


Построение экспоненциального тренда на графике

```
plot(s$год, s$s, col="orange2", type="l", xlab="",  
     ylab="Численность населения России, млн. чел.", main =  
     'Экспоненциальный тренд')  
points(s[,1], predict(lm(s$s~exp(s$t))), pch=20, col="red", type="p")
```

```
ss = lm(s$s ~ exp(s$t))  
summary(ss)
```

```
Call:  
lm(formula = s$s ~ exp(s$t))  
  
Residuals:  
    Min       1Q   Median       3Q      Max   
-146.566  -92.197   -2.925    79.867   161.817  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)  6.660e+02  2.463e+01  27.037 5.01e-16 ***  
exp(s$t)     -5.318e-07  2.111e-07  -2.519  0.0215 *    
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 104 on 18 degrees of freedom  
Multiple R-squared:  0.2606,    Adjusted R-squared:  0.2195   
F-statistic: 6.344 on 1 and 18 DF,  p-value: 0.02145
```

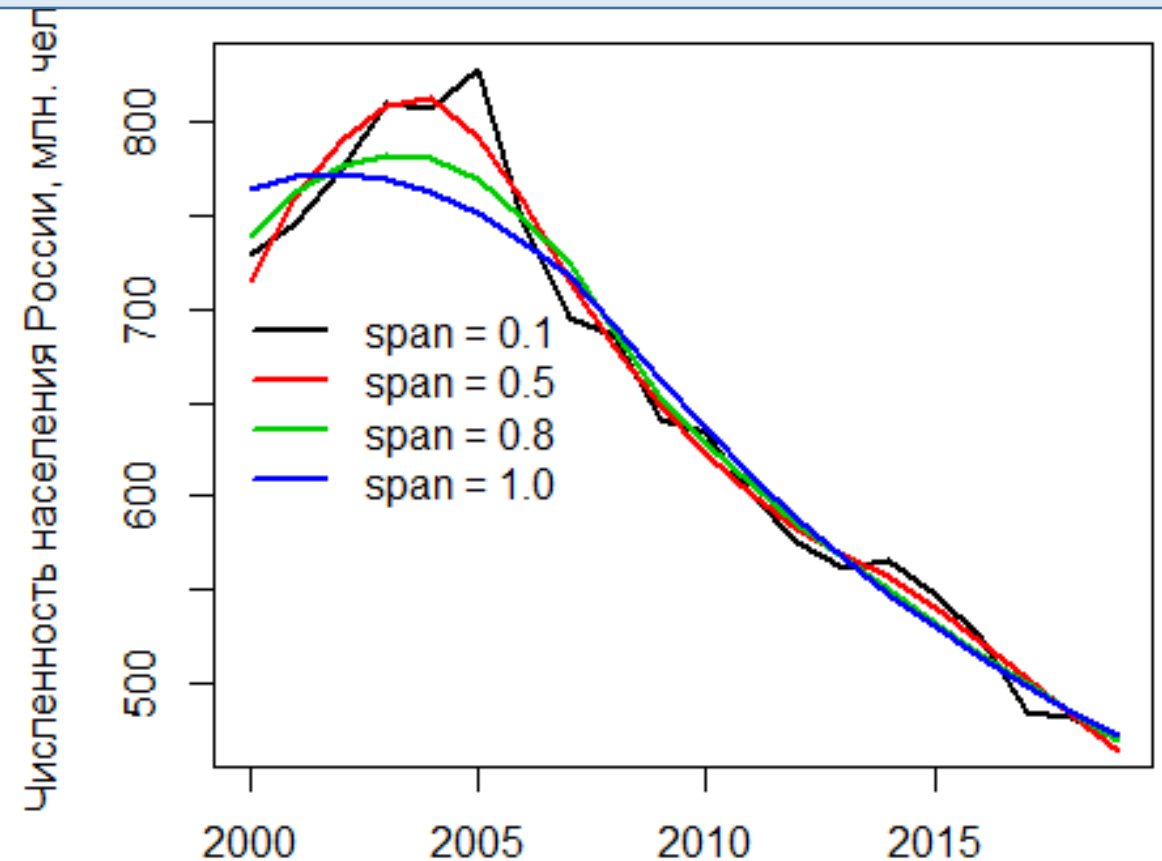


Программа подбора в R кривой для аппроксимации численности населения России

```
# Создаем список значений параметра сглаживания
spanlist <- c(0.15, 0.50, 0.8, 1.0)
plot(s$год, s$s, col="red", type="l", xlab="",
     ylab="Численность населения России, млн. чел.")
for (i in 1:length(spanlist)) {
  T.loess <- loess( s$s~ s$год, span=spanlist[i])
  T.predict <- predict(T.loess)
  lines(s$год,T.predict,col=i, lwd=2)
}
legend ("left",c(paste("span =", formatC(spanlist,digits=1,
                                           format="f"))), col=1:length(spanlist),
       lwd=2, bty="n")
```

Аппроксимация динамики показателя с использованием локальной регрессии, реализованной в R функцией **loess()**

Чтобы подобрать адекватную степень сглаживания кривой, выполним построение серии моделей, варьируя величину span



2. Модели сезонности: мультипликативная и аддитивная

Сезонные колебания показателя (сезонность в данных) – устойчивые внутри годичные колебания, то есть когда из года в год в одни и те же периоды внутри года уровень повышается, а в другие понижается.

ИНДЕКСЫ СЕЗОННОСТИ

Индексы сезонности – показатели измеряющие сезонность.

Их расчет зависит от тенденции показателя:

1. Если годовой уровень из года в год существенно не изменяется (нет тренда) (**на основе средних уровней**);
2. Если годовой уровень из года в год существенно изменяется (существует тренд) (**на основе тренда**).

РАСЧЕТ ИНДЕКСОВ СЕЗОННОСТИ НА ОСНОВЕ СРЕДНИХ УРОВНЕЙ

- 1 этап – за все годы исследуемого периода рассчитать средние уровни по одноименным внутригодовым периодам \bar{Y}_k
- 2 этап – рассчитать средний уровень ряда динамики за весь период \bar{Y}
- 3 этап – рассчитать столько индексов сезонности, сколько внутригодовых периодов в году (кварталов – 4; месяцев – 12 и т.д.) по формуле:

$$I_s^k = \frac{\bar{Y}_k}{\bar{Y}}$$

РАСЧЕТ ИНДЕКСОВ СЕЗОННОСТИ НА ОСНОВЕ ТРЕНДА

- 1 этап - за все годы исследуемого периода рассчитать средние уровни по одноименным внутригодовым периодам \bar{Y}_k
- 2 этап – выделить скользящее среднее или тренд \hat{Y}_t
- 3 этап – найти средние уровни по сглаженным или выровненным одноименным периодам $\hat{\bar{Y}}_k$
- 4 этап - найти индексы сезонности:

$$I_s^k = \frac{\bar{Y}_k}{\hat{\bar{Y}}_k}$$

СТЕПЕНЬ ВЛИЯНИЯ СЕЗОННОСТИ

Для сопоставления сезонности по нескольким рядам динамики исчисляют ошибку сезонности по (m – число внутригодовых периодов анализа):

$$\sigma_s = \sqrt{\frac{\sum (I_s - 1)^2}{m}}$$

Чем меньше данный показатель, тем меньше влияет сезонность

УЧЕТ СЕЗОННОСТИ В ПРОГНОЗЕ

Для того, чтобы прогноз был более точным, нужно учесть сезонность, для чего уровни прогноза умножаются на индексы сезонности внутригодовому соответственно периоду.

$$Y_k^{прогноз} = \hat{Y}_k \cdot I_s^k$$

Функция `decompose(Y, type="additive", filter = NULL)` из пакета `stats`

↑
Y - переменная в виде
динамического ряда

↑
Тип модели

↑
Выбор метода
сглаживания,
NULL – метод
скользящей средней

пусть `d = decompose(Y, type="additive", filter = NULL)`, то
`d$trend` – значения тренда, вычисленные по методу скользящей
средней, `d$figure` – значения индексов сезонности, `d$random` -
остатки

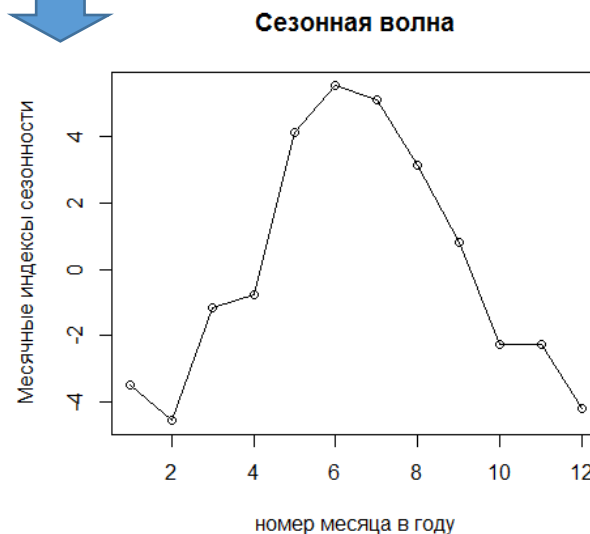
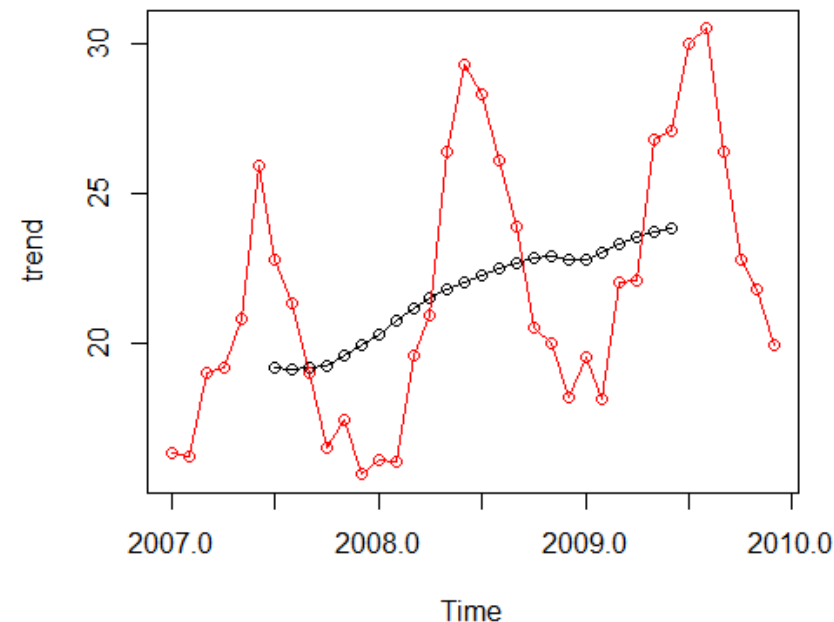
Пример

	A	B	C	D
5	Апреля	4	19,20	
6	Май	5	20,80	
7	Июнь	6	25,90	
8	Июль	7	22,80	
9	Август	8	21,30	
10	Сентября	9	19,00	
11	Октября	10	16,50	
12	Ноября	11	17,40	
13	Декабря	12	15,60	
14	Января	13	16,10	
15	Февраля	14	16,00	
16	Март	15	19,60	
17	Апреля	16	20,90	
18	Май	17	26,40	
19	Июнь	18	29,30	
20	Июль	19	28,30	
21	Август	20	26,10	
22	Сентября	21	23,90	
23	Октября	22	20,50	
24	Ноября	23	20,00	
25	Декабря	24	18,20	
26	Января	25	19,50	
27	Февраля	26	18,10	
28	Март	27	22,00	
29	Апреля	28	22,10	
30	Май	29	26,80	
31	Июнь	30	27,10	
32	Июль	31	30,00	
33	Август	32	30,50	
34	Сентября	33	26,40	
35	Октября	34	22,80	
36	Ноября	35	21,30	
37	Декабря	36	19,90	
38				

```
#сезонность
types = c("text", rep("numeric", 2))
yt <- as.data.frame(read_excel("c:/Users/компьютер/Documents/diny.xlsx", 1,
                              col_types = types))

Yt=yt[,3]
y=ts(Yt,start=c(2007,1), end=c(2009,12), frequency = 12)
y
t=c(1:length(y))
library(stats)
d=decompose(y, type = "additive", filter = NULL)
s=d$figure
Tr=d$trend
e=d$random
ts.plot(Tr,y, ylab="trend", col=1:2, type="o")
```

```
#сезонная волна
tt=c(1:length(s))
plot ( tt, s, col=1, type="o", main="Сезонная волна",
      ylab="Месячные индексы сезонности",
      xlab="номер месяца в году")
```



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	16.3	16.2	19.0	19.2	20.8	25.9	22.8	21.3	19.0	16.5	17.4	15.6
2008	16.1	16.0	19.6	20.9	26.4	29.3	28.3	26.1	23.9	20.5	20.0	18.2
2009	19.5	18.1	22.0	22.1	26.8	27.1	30.0	30.5	26.4	22.8	21.8	19.9

3. Модели тренда и сезонности

в нашем примере

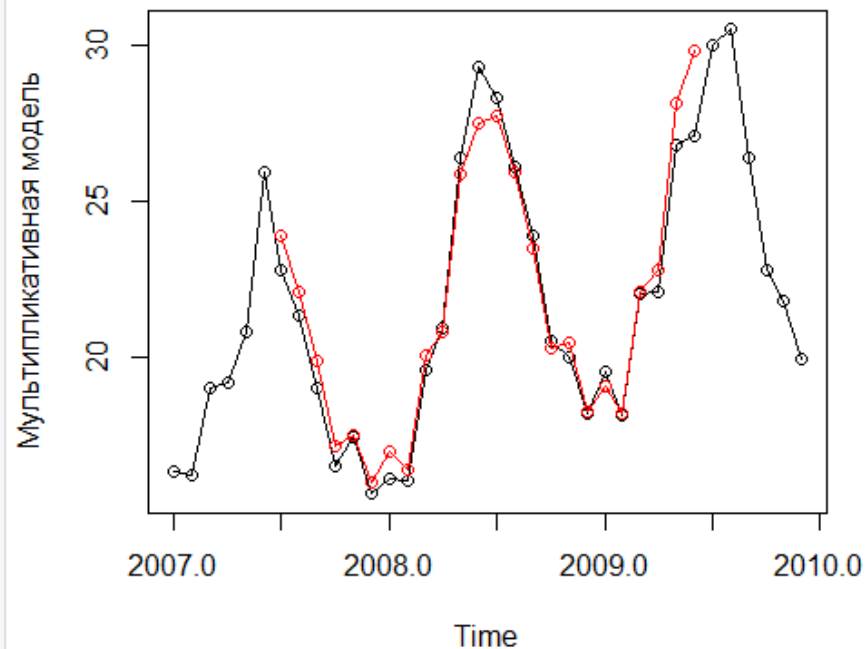
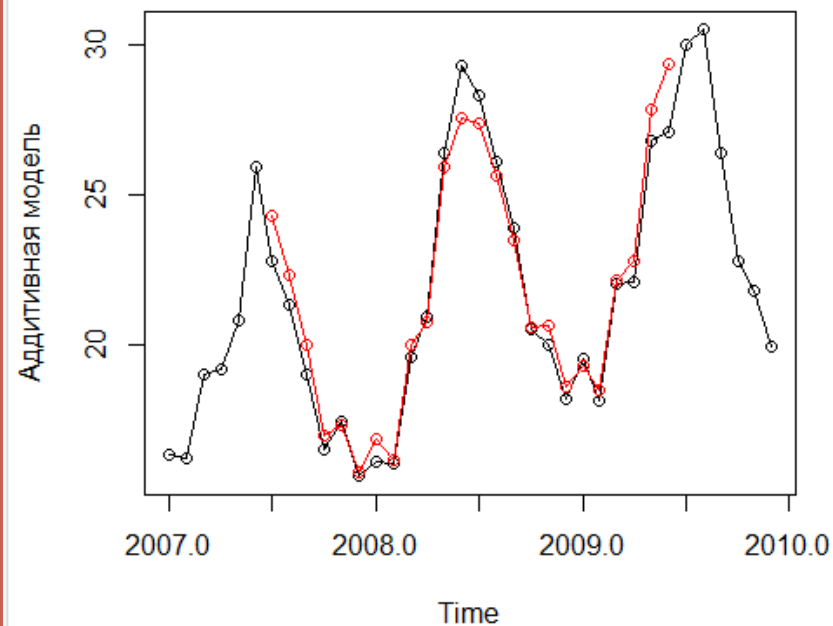
$yf = Tr + s$

`ts.plot(y, yf, ylab="Аддитивная модель", col=1:2, type="o")`

```
d=decompose(y, type = "multiplicative" )
s=d$figure
s
#[1] 0.8354833 0.7891474 0.9477889 0.9684083 1.1872081 1.2508308 1.2476421 1.1522889 1.0369078
#[10] 0.8895802 0.8937466 0.8009676
```

```
Tr=d$trend
e=d$random
ts.plot(Tr,y, ylab="trend", col=1:2, type="o")
```

```
#сезонная волна
tt=c(1:length(s))
plot ( tt, s, col=1, type="o", main="Сезонная волна",
      ylab="месячные индексы сезонности", xlab="номер месяца в году")
yf=Tr*s
ts.plot(y,yf, ylab="мультипликативная модель", col=1:2, type="o")
```



В R строятся **аддитивные регрессионные модели** (Generalized Additive Models, GAM) следующего вида:

$$y(t)=g(t)+s(t)+h(t)+\epsilon t,$$

где $g(t)$ и $s(t)$ — функции, аппроксимирующие тренд ряда и сезонные колебания (например, годовые, недельные и т.п.) соответственно, $h(t)$ — функция, отражающая эффекты праздников и других влиятельных событий, а ϵt — нормально распределённые случайные возмущения (остатки).

В R строятся **аддитивные регрессионные модели** (Generalized Additive Models, GAM) следующего вида:

$$y(t)=g(t)+s(t)+h(t)+\epsilon t,$$

где $g(t)$ и $s(t)$ — функции, аппроксимирующие тренд ряда и сезонные колебания (например, годовые, недельные и т.п.) соответственно, $h(t)$ — функция, отражающая эффекты праздников и других влиятельных событий, а ϵt — нормально распределённые случайные возмущения (остатки).

Панельные данные

- Термин "панельные данные" (panel data) пришёл из обследований индивидов, и в этом контексте "панель" представляла собой группу индивидов, за которыми регулярно осуществляли наблюдения в течение определённого периода времени.
- В настоящее время методы анализа панельных данных получили большое распространение, и понимание панельных данных стало намного шире. Наряду с термином "панельные данные" иногда также используется термин "лонгитюдные данные" (longitudinal data).

Историческая справка

- Впервые панельные данные начали формироваться в США в 1960-х гг. Среди наиболее известных баз панельных данных США можно выделить PSID и NLS.
- Панельное исследование динамики доходов (The US Panel Study of Income Dynamics (PSID), [psidonline.isr.umich.edu /](http://psidonline.isr.umich.edu/)) – база панельных данных по американским домохозяйствам, собираемая Институтом социальных исследований Мичиганского университета. База PSID появилась в 1968 г. и включала данные по 4800 семьям. В настоящее время она охватывает около 9000 американских семей. Данные содержат более 5000 переменных по экономике, демографии, здоровью и социальному поведению.

Европейские панельные данные

- стали появляться только в 1980-х гг. Так, например первая волна панельного обследования Немецкой социально-экономической панели (Soziooekonomisches Panel (SOEP), diw.de/soep), формируемой Немецким институтом экономических исследований (Deutsches Institut für Wirtschaftsforschung (DIW), Berlin), состоялась в 1984 г. и охватила более 5000 западногерманских домохозяйств.
- В настоящее время это обследование содержит данные около 11 000 домохозяйств, которые включают в себя демографические переменные, зарплату, доход, выплату пособий, уровень удовлетворенности различными аспектами жизни, надежды и страхи, политическую активность и т.д.

Британские панельные данные

- С 1991 г. Институтом социальных и экономических исследований Эссекского университета проводится панельное исследование британских домохозяйств (The British Household Panel Survey (BHPS), [iser.essex.ac.uk/survey / bhps](http://iser.essex.ac.uk/survey/bhps)). Это национальная репрезентативная выборка 5500 домохозяйств и 10 300 индивидов, выбранных из 250 районов Великобритании. Эти данные отражают демографические характеристики домохозяйств, рынок труда, здоровье, образование, жилищные условия, потребление, доход и т.д. В 1994– 2001 гг. при содействии Евростата проводилось Европейское панельное обследование домохозяйств (The European Community Household Panel (ECHP), [epunet.essex.ac.uk / echpphp](http://epunet.essex.ac.uk/echpphp))

Российские панельные данные

- В России панельные обследования стали проводиться в 1990-х г. Наиболее известной базой панельных данных является РМЭЗ – Российский мониторинг экономического положения и здоровья населения (Russia Longitudinal Monitoring Survey (RLMS), [cpc.unc.edu /projects / rims](http://cpc.unc.edu/projects/rims)).
- РМЭЗ представляет собой серию общенациональных репрезентативных опросов домохозяйств и индивидов, проводившихся в России с 1992 г.

Панельные данные

- состоят из **повторных наблюдений** одних и тех же выборочных единиц, которые осуществляются в последовательные периоды времени.
- В качестве объектов наблюдения могут выступать индивиды, домашние хозяйства, фирмы, страны и т.д.
- Примером панельных данных могут быть ежегодные обследования одних и тех же домашних хозяйств или индивидов (например, для определения изменения их благосостояния), ежеквартальные данные об экономической деятельности отдельных компаний, ежегодные социально-экономические показатели для регионов одной страны или для группы стран и т.д.

Панельные данные

- совмещают в себе как пространственные данные, так и временные ряды и сочетают достоинства каждого из этих видов данных. Это позволяет строить **более адекватные** и содержательные модели для изучения истинной причинно-следственной связи между различными переменными, что представляется невозможным в рамках только временных или только пространственных данных.

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.

Преимущества использования панельных данных

- 2. 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.

Преимущества использования панельных данных

- 3. Панельные данные предоставляют возможность **изучать динамику изменений индивидуальных характеристик единиц совокупности**.
- Панельные данные хорошо подходят для изучения перемены работы, периода безработицы, изменений в доходах, для исследования длительности пребывания в определённом экономическом состоянии, например в бедности или в качестве безработного, а также могут помочь изучить скорость приспособления индивидов к изменениям в экономической политике.

Преимущества использования панельных данных

- 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 счёт 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.

Преимущества использования панельных данных

- 5. Панельные данные позволяют конструировать и тестировать более сложные **поведенческие модели**, чем пространственные данные и временные ряды в отдельности. Например, техническая эффективность лучше изучается и моделируется с панельными данными. Также в панелях может быть наложено меньше ограничений на модели распределенного лага, которые обычно рассматриваются во временных рядах.

Преимущества использования панельных данных

- 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.

Преимущества использования панельных данных

- 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.

Недостатки панельных данных

- **проблема покрытия**, т.е. неполный учёт интересующей совокупности;
- **отсутствие отклика**, которое может быть связано как с отсутствием взаимодействия с респондентом, так и с ошибкой интервьюера, искажения, связанные с ошибками измерения, которые могут возникнуть по причине неправильного ответа из-за неясной формулировки вопроса, ошибок памяти, намеренного искажения ответа (престижное смещение), неподходящих информантов, ошибочной записи ответов и эффектов интервьюера.

Формализация

- Имеется множество объектов (индивидуумы, домашние хозяйства, фирмы, регионы, страны и т.п.), занумерованных индексами $i=1, \dots, n$.
- Они наблюдаются в моменты времени $t=1, \dots, T$. Каждый рассматриваемый объект характеризуется k переменными (признаками):

$$x_{it} = (x_{it}^1, \dots, x_{it}^k) \in \mathbb{R}^k.$$

Пример

- Объекты – коммерческие фирмы;
- x_i – оборот, прибыль, число сотрудников, отрасль; y_i – рыночная стоимость.

Обозначения

Введем обозначения:

- x_{it} – набор независимых переменных (вектор размерности k)
- y_{it} – зависимая переменная для экономической единицы i в момент времени t
- ϵ_{it} – соответствующая ошибка.
- Обозначим также:

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad X_i = \begin{bmatrix} x'_{i1} \\ \vdots \\ x'_{iT} \end{bmatrix}, \quad \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iT} \end{bmatrix}.$$

- Введем также «объединенные» наблюдения и ошибки:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Здесь y , ϵ – $nT \times 1$ векторы, X – $nT \times k$ матрица.

Частные случаи

- общее число наблюдений будет nT .
- 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 $T > 1$.

Виды панелей

Сбалансированные панели

- для каждой пространственной единицы имеется одинаковое число наблюдений по всем периодам времени.
- Несбалансированные панели

Основные модели анализа панельных данных

1. Объединенная модель панельных данных (Pooled model)
2. Модель панельных данных с фиксированными эффектами (Fixed effect model)
3. Модель панельных данных со случайными эффектами (Random effect model)

Описание объединённой модели

Это обычная линейная модель регрессии

Считается, что зависимая переменная линейно зависит от всех переменных в тот же момент времени.

$$y_{it} = x'_{it} \cdot \beta + \mu = \sum_{j=1}^k x_{it}^j \cdot \beta_j + \mu$$

или в матричной форме

$$y = X \cdot \beta + \mu,$$

которая, по существу, не учитывает панельную структуру данных. (Здесь β – неизвестный вектор размера $k \times 1$.)

pooled model

- Для настройки параметров можно использовать метод наименьших квадратов

$$\sum_{i=1}^n \sum_{t=1}^T (\hat{y}_{it} - y_{it})^2 \rightarrow \min_{\beta, \mu}.$$

$$\beta \in \mathbb{R}^k, \mu \in \mathbb{R}.$$

fixed effect model

- Модель панельных данных с фиксированными эффектами опирается на структуру панельных данных, что позволяет учитывать неизмеримые индивидуальные различия объектов. Эти отличия называются **эффектами**.
- В данной модели эффекты интерпретируются как мешающий параметр, и оценивание направлено на то, чтобы их исключить.

Введем обозначения:

- $i = 1, \dots, n$ – номера объектов, $t = 1, \dots, T$ – моменты времени, k – число признаков.
- x_{it} – набор независимых переменных (вектор размерности k)
- y_{it} – зависимая переменная для экономической единицы i в момент времени t
- ε_{it} – соответствующая ошибка.
- Обозначим также:

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}, X_i = \begin{bmatrix} x'_{i1} \\ \vdots \\ x'_{iT} \end{bmatrix}, \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix}.$$

- Введем также «объединенные» наблюдения и ошибки:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Здесь y, ε – $nT \times 1$ векторы, X – $nT \times k$ матрица.

fixed effect model

- модель панельных данных с фиксированными эффектами описывается уравнением:

$$y_{it} = \alpha_i + x'_{it} \cdot \beta + \epsilon_{it}.$$

Величина α_i выражает индивидуальный эффект объекта i , не зависящий от времени t , при этом регрессоры x_{it} не содержат константу.

Параметры модели: $\beta \in \mathbb{R}^k, \alpha_i \in \mathbb{R} (i = 1, \dots, n)$.

Основные предположения

Предположим, что выполнены следующие условия:

1. ошибки ϵ_{it} некоррелированы между собой по i и t , $E(\epsilon_{it}) = 0$, $V(\epsilon_{it}) = \sigma_\epsilon^2$;
2. ошибки ϵ_{it} некоррелированы с регрессорами x_{js} при всех i, j, t, s .

Понижение размерности. Исключение эффектов.

Для панельных данных типична ситуация, когда число объектов n достаточно велико. Поэтому, применяя непосредственно **метод наименьших квадратов** к уравнению , при оценивании параметров можно столкнуться с вычислительными проблемами. Их можно преодолеть, исключая из рассмотрения индивидуальные эффекты α_i . При этом мы *понижаем размерность задачи с $(n+k)$ до k* .

Наиболее простой способ – переход в уравнении к средним по времени величинам:

$$\overline{y_i} = \alpha_i + \overline{x_i'} \cdot \beta + \overline{\epsilon_i},$$

где $\overline{y_i} = \frac{1}{T} \sum_{t=1}^T y_{it}$, $\overline{x_i} = \frac{1}{T} \sum_{t=1}^T x_{it}$, $\overline{\epsilon_i} = \frac{1}{T} \sum_{t=1}^T \epsilon_{it}$.

Вычитая почленно (2) из (1), получаем:

$$y_{it} - \overline{y_i} = (x_{it} - \overline{x_i})' \cdot \beta + \epsilon_{it} - \overline{\epsilon_i}.$$

Данная модель уже не зависит от эффектов α_i . По существу, это уравнение , записанное в отклонениях от индивидуальных средних по времени.

Наиболее простой способ – переход в уравнении к средним по времени величинам:

$$\overline{y_i} = \alpha_i + \overline{x_i'} \cdot \beta + \overline{\varepsilon_i}, \quad (1)$$

где $\overline{y_i} = \frac{1}{T} \sum_{t=1}^T y_{it}$, $\overline{x_i} = \frac{1}{T} \sum_{t=1}^T x_{it}$, $\overline{\varepsilon_i} = \frac{1}{T} \sum_{t=1}^T \varepsilon_{it}$. (2)

Вычитая почленно (2) из (1), получаем:

$$y_{it} - \overline{y_i} = (x_{it} - \overline{x_i})' \cdot \beta + \varepsilon_{it} - \overline{\varepsilon_i}. \quad (3)$$

Данная модель уже не зависит от эффектов α_i . По существу, это уравнение ,
записанное в отклонениях от индивидуальных средних по времени.

Оценка параметров модели

Применяя обычный [метод наименьших квадратов](#) к уравнению (3), мы получим оценки

$$\widehat{\beta} = \left(\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i) \cdot (x_{it} - \bar{x}_i)' \right)^{-1} \cdot \sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i) \cdot (y_{it} - \bar{y}_i) \quad (4)$$

Эти оценки называются **внутригрупповыми оценками** (**within estimator**) или **оценками с фиксированным эффектом** (**fixed effect estimator**).

Условия 1)-2), наложенные на модель, гарантируют [несмещённость](#) и [состоятельность](#) оценок с фиксированным эффектом.

В качестве оценок индивидуальных эффектов можно взять

$$\widehat{\alpha}_i = \bar{y}_i - \bar{x}_i' \cdot \widehat{\beta}, \quad i = 1, \dots, n.$$

Эти оценки являются [несмещёнными](#) и [состоятельными](#) для фиксированного n при $t \rightarrow \infty$.

Из формулы (4) вытекает выражение для **матрицы ковариации** оценки $\widehat{\beta}$:

$$V(\widehat{\beta}) = \sigma_{\varepsilon}^2 \left(\sum_{i=1}^n \sum_{t=1}^T (x_{it} - \bar{x}_i) \cdot (x_{it} - \bar{x}_i)' \right)^{-1}.$$

Как и в обычной линейной модели, в качестве оценки дисперсии σ_{ε}^2 можно взять [сумму квадратов остатков регрессии](#), деленную на число степеней свободы:

$$\widehat{\sigma}_{\varepsilon}^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)' \widehat{\beta})^2}{nT - n - k}.$$

При достаточно слабых условиях регулярности оценки с фиксированным эффектом являются *асимптотически нормальными* (при $n \rightarrow \infty$ или при $T \rightarrow \infty$), поэтому можно пользоваться стандартными процедурами (t -тесты, F -тесты) для проверки гипотез относительно параметров β .

Недостатки модели панельных данных с фиксированными эффектами

В панельных данных среди независимых переменных x_{it} могут быть такие, которые не меняются во времени для каждого объекта. Например, при анализе заработной платы в число факторов часто включают пол или расовую принадлежность. Модель с фиксированным эффектом не позволяет идентифицировать соответствующие таким переменным коэффициенты. Формально это объясняется тем, что в уравнении (3) один или несколько регрессоров равны нулю, и, следовательно, метод наименьших квадратов применять нельзя.

time-varying model for panel data

- Модель панельных данных с временными эффектами (time-varying model for panel data) опирается на структуру панельных данных, что позволяет учитывать неизмеримые индивидуальные различия объектов. Эти отличия называются эффектами.
- В данной модели **эффекты объектов могут изменяться в каждый момент времени.**

Обозначения

Введем обозначения:

- $i = 1, \dots, n$ – номера объектов, $t = 1, \dots, T$ – моменты времени, k – число признаков.

Для каждого объекта в каждый момент времени известны:

- x_{it} – набор независимых переменных (вектор размерности k)
- y_{it} – зависимая переменная для экономической единицы i в момент времени t

Описание модели панельных данных с временными эффектами

В введенных обозначениях модель панельных данных с временными эффектами описывается уравнением

$$\hat{y}_{it} = \alpha_i + \gamma_t + x'_{it}\beta \quad (1)$$

Здесь \hat{y}_{it} модельное значение зависимой переменной, соответствующее y_{it} . Величина α_i выражает индивидуальный эффект объекта i , не зависящий от времени t . Величина γ_t выражает зависимость индивидуального эффекта объекта i от времени t (будем считать, что всегда $\gamma_1 = 1$). При этом регрессоры x_{it} не содержат константу.

Параметры модели: $\beta \in \mathbb{R}^k, \alpha_i \in \mathbb{R} (i = 1, \dots, n), \gamma_t \in \mathbb{R} (t = 2, \dots, T)$.

Понижение размерности. Исключение эффектов

Для панельных данных типична ситуация, когда число объектов n достаточно велико. Поэтому, применяя непосредственно **метод наименьших квадратов** к уравнению (1), при оценивании параметров можно столкнуться с вычислительными проблемами. Их можно преодолеть, исключая из рассмотрения индивидуальные эффекты $\alpha_i \cdot \gamma_t$. При этом мы *понижаем размерность задачи с $(n+k+T-1)$ до k* .

Наиболее простой способ – переход в уравнении (1) к средним величинам по времени и по множеству объектов :

$$\overline{y}_i = \alpha_i + \overline{\gamma} + \overline{x}_i' \cdot \beta, \quad (2)$$

$$\text{где } \overline{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}, \quad \overline{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}, \quad \overline{\gamma} = \frac{1}{T} \sum_{t=1}^T \gamma_t;$$

$$\overline{y}_t = \overline{\alpha} + \gamma_t + \overline{x}_t' \cdot \beta, \quad (3)$$

$$\text{где } \overline{y}_t = \frac{1}{n} \sum_{i=1}^n y_{it}, \quad \overline{x}_t = \frac{1}{n} \sum_{i=1}^n x_{it}, \quad \overline{\alpha} = \frac{1}{n} \sum_{i=1}^n \alpha_i$$

$$\overline{y} = \overline{\alpha} + \overline{\gamma} + \overline{x}' \cdot \beta, \quad (4)$$

$$\text{где } \overline{y} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T y_{it}, \quad \overline{x} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T x_{it},$$

Из (1),(2),(3),(4) получаем:

$$\mathcal{Y}_{it} - \overline{y}_i - \overline{y}_t + \overline{y} = (x_{it} - \overline{x}_i - \overline{x}_t + \overline{x})' \cdot \beta. \quad (5)$$

Данная модель уже не зависит от эффектов $\alpha_i \cdot \gamma_t$.

Заметим, что если в рассмотренной модели все коэффициенты $\gamma_t = 1 (t = 1, \dots, T)$, то получим **Модель панельных данных с фиксированными эффектами**. На практике, чтобы понять, какая из этих двух моделей адекватнее, можно проверить **гипотезу** $H_0: \gamma_1 = \dots = \gamma_T = 1$. Обычно для этого используют **F-тест**.

Проблемы

Для устойчивости данной модели необходимо, чтобы значения γ_t изменялись плавно. Для этого используют регуляризацию:

$$\sum_{i=1}^n \sum_{t=1}^T (\hat{y}_{it} - y_{it})^2 + \lambda \sum_{t=2}^T (\gamma_t - \gamma_{t-1})^2 \longrightarrow \min_{\alpha, \beta, \gamma}$$

Тогда, чем больше значение λ , тем более гладко изменяется γ_t . Для выбора значения λ можно использовать метод **скользящего контроля**.

Ротационная панель

- модель [панельных данных](#), в которой у каждого объекта есть своё время жизни. Данную модель удобно использовать, если информация об объекте доступна лишь в некоторые моменты времени.

Обозначения

Введем обозначения:

- x_{it} – набор независимых переменных ($i = 1, \dots, k$ $t = 1, \dots, T$),
- y_{it} – зависимая переменная.
- Обозначим также: $\Omega_i = [T_{1i}, T_{2i}]$ - время жизни i го объекта.

Описание ротационной модели

Построим обычную модель **линейной регрессии**, с учетом того, что в данных есть пропуски:

$$\hat{y}_{it} = x'_{it} \cdot \beta + \mu, \quad t \in \Omega_i, \quad i = 1, \dots, k$$

Параметры модели: $\beta \in \mathbb{R}^k, \mu \in \mathbb{R}$.

Для настройки параметров можно использовать **метод наименьших квадратов** с небольшой модификацией:

$$\sum_{i=1}^n \sum_{t=T_{1i}}^{T_{2i}} (\hat{y}_{it} - y_{it})^2 \rightarrow \min_{\beta, \mu}.$$

Пример построения моделей по панельным данным в R

```
#делаем панельные данные , чтобы данные об одинаковых регионах за два
#периода были подряд
#создаем 2 панели
panel1=alldata[,2]
panel1=as.data.frame(panel1)
panel1[,2:12]=y[,c(1:11)]
panel2=alldata[,2]
panel2=as.data.frame(panel2)
panel2[,2:12]=y[,c(12:22)]
#называем столбцы одинаково
colnames(panel1)[12]="y"
colnames(panel2)[12]="y"
colnames(panel2)[1]="регион"
colnames(panel1)[1]="регион"
colnames(panel2)[2]="x1"
colnames(panel1)[2]="x1"
colnames(panel2)[11]="x10"
colnames(panel2)[11]="x10"
#объединяем панели в одну
panel3<- rbind(panel1,panel2)
#сортируем по регионам
panel3=panel3[order(panel3$регион), ]
```

```
#подключаем пакет для построения моделей панельных
данных
install.packages("plm")
library("plm")
#делаем модель объединенной модели регрессии
#спецификация модели
fm=panel3$y~panel3$x1+panel3$x2+panel3$x3+panel3$x4+
panel3$x5+panel3$x6+panel3$x7+panel3$x8+panel3$x9+pa
nel3$x10

pool=plm(fm,data=panel3,index=c("регион"),
model="pooling", effect="individual")
summary(pool)
```

```

> pool=plm(fm,data=panel3,index=c("регион"), model="pooling", effect="individual")
> summary(pool)
Pooling Model

Call:
plm(formula = fm, data = panel3, effect = "individual",
     model = "pooling", index = c("регион"))

Balanced Panel: n = 87, T = 2, N = 174

Residuals:
    Min.    1st Qu.    Median    3rd Qu.    Max.
-0.2728305 -0.0210618 -0.0017216  0.0229469  0.2920806

Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
(Intercept)  2.29133799  0.85928685   2.6666 0.008436 **
panel3$x1    -0.01934199  0.02314969  -0.8355 0.404649
panel3$x2     0.08570618  0.08074488   1.0614 0.290058
panel3$x3    -0.10264406  0.17547060  -0.5850 0.559380
panel3$x4     0.09067710  0.04399035   2.0613 0.040863 *
panel3$x5    -0.00079716  0.00023923  -3.3323 0.001066 **
panel3$x6    -0.03892990  0.02722443  -1.4300 0.154642
panel3$x7     0.13780114  0.06858836   2.0091 0.046177 *
panel3$x8     0.03466352  0.02055388   1.6865 0.093618 .
panel3$x9    -0.00051678  0.00074708  -0.6917 0.490091
panel3$x10   -1.44613757  0.84469453  -1.7120 0.088794 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.58496
Residual Sum of Squares: 0.44318
R-Squared:               0.24238
Adj. R-Squared:          0.1959
F-statistic: 5.21469 on 10 and 163 DF, p-value: 1.2539e-06

```


#модель с фиксированными эффектами

```
fm=panel3$y~panel3$x1+panel3$x2+panel3$x3+panel3$x4+panel3$x5+p  
anel3$x6+panel3$x7+panel3$x8+panel3$x9+panel3$x10
```

```
summary(fm)
```

```
fe=plm(fm,data=panel3,index=c("регион"), model="within",  
effect="individual")
```

```
summary(fe)
```

```
summary(fixef(fe))
```

```
> fe=plm(fm,data=panel3,index=c("регион"), model="within", effect="individual")
```

```
> summary(fe)
```

Oneway (individual) effect within Model

Call:

```
plm(formula = fm, data = panel3, effect = "individual",  
     model = "within", index = c("регион"))
```

Balanced Panel: n = 87, T = 2, N = 174

Residuals:

	Min.	1st Qu.	Median	3rd Qu.	Max.
	-1.3982e-01	-1.4744e-02	7.4322e-17	1.4744e-02	1.3982e-01

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
panel3\$x1	-0.03443115	0.03814616	-0.9026	0.3695
panel3\$x2	0.17542407	0.12384972	1.4164	0.1607
panel3\$x3	-0.03962732	0.27187840	-0.1458	0.8845
panel3\$x4	0.10328725	0.06605678	1.5636	0.1220
panel3\$x5	-0.00045973	0.00035286	-1.3029	0.1965
panel3\$x6	0.02727210	0.03622880	0.7528	0.4539
panel3\$x7	0.00743075	0.10033580	0.0741	0.9412
panel3\$x8	0.03586259	0.02672085	1.3421	0.1835
panel3\$x9	0.00013469	0.00106581	0.1264	0.8998
panel3\$x10	-1.41872630	1.34013962	-1.0586	0.2931

Total Sum of Squares: 0.32296

Residual Sum of Squares: 0.21464

R-Squared: 0.33542

Adj. R-Squared: -0.49315

F-statistic: 3.88627 on 10 and 77 DF, p-value: 0.00027019

```
> summary(fixef(fe))
```

	Estimate	Std. Error	t-value	Pr(> t)
Алтайский край	2.1578	1.3993	1.5420	0.12716
Амурская область	2.1946	1.3970	1.5709	0.12030
Архангельская область	2.1641	1.3983	1.5477	0.12580
Архангельская область без авт. округа	2.1690	1.3983	1.5512	0.12495
Астраханская область	2.1546	1.3917	1.5482	0.12568
Белгородская область	2.1525	1.3964	1.5415	0.12730
Брянская область	2.1584	1.3950	1.5473	0.12589
Владимирская область	2.1484	1.3934	1.5419	0.12719
Волгоградская область	2.1255	1.3970	1.5215	0.13223
Вологодская область	2.1565	1.3963	1.5444	0.12660
Воронежская область	2.1286	1.3959	1.5249	0.13138
г. Москва	2.1333	1.3906	1.5340	0.12912
г. Санкт-Петербург	2.1241	1.3916	1.5264	0.13100
г. Севастополь	2.0702	1.4003	1.4784	0.14338
Еврейская авт. область	2.0217	1.3867	1.4579	0.14893
Забайкальский край	2.1733	1.3934	1.5598	0.12291
Ивановская область	2.1621	1.3955	1.5493	0.12541
Иркутская область	2.1332	1.3872	1.5378	0.12820
Кабардино-Балкарская Республика	2.1272	1.3940	1.5260	0.13112
Калининградская область	2.1431	1.3909	1.5408	0.12746
Калужская область	2.1369	1.3934	1.5336	0.12922
Камчатский край	2.1084	1.3966	1.5097	0.13522
Карачаево-Черкесская Республика	2.0736	1.3988	1.4824	0.14231
Кемеровская область	2.1230	1.3942	1.5228	0.13191
Кировская область	2.1532	1.4020	1.5358	0.12868
Костромская область	2.1154	1.4006	1.5103	0.13505
Краснодарский край	2.0994	1.3988	1.5009	0.13747
Красноярский край	2.1780	1.3945	1.5618	0.12243
Курганская область	2.1547	1.3970	1.5424	0.12708
Курская область	2.1396	1.3979	1.5306	0.12996
Ленинградская область	2.1527	1.4275	1.5080	0.13565
Магнитогорская область	2.1330	1.3960	1.5280	0.13000

#модель со случайными эффектами

```
fm=panel3$y~panel3$x1+panel3$x2+panel3$x3+panel3$x4+panel3$x5+panel3$x6+panel3$x7+panel3$x8+panel3$x9+panel3$x10
```

```
fb=plm(fm,data=panel3,index=c("регион"),  
model="between", effect="individual")
```

```
summary(fb)
```

```

> fb=plm(fm,data=panel3,index=c("регион"), model="between", effect="individual")
> summary(fb)
Oneway (individual) effect Between Model

Call:
plm(formula = fm, data = panel3, effect = "individual",
     model = "between", index = c("регион"))

Balanced Panel: n = 87, T = 2, N = 174
Observations used in estimation: 87

Residuals:
      Min.      1st Qu.      Median      3rd Qu.      Max.
-0.12386612 -0.01694934  0.00025941  0.01667673  0.17417656

Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
(Intercept)  2.2634713   1.1052554   2.0479 0.044023 *
panel3$x1    -0.0177441   0.0302329  -0.5869 0.559000
panel3$x2     0.0459406   0.1123667   0.4088 0.683803
panel3$x3    -0.3952515   0.2372544  -1.6659 0.099842 .
panel3$x4     0.0356182   0.0702129   0.5073 0.613421
panel3$x5    -0.0021769   0.0029599  -0.7355 0.464320
panel3$x6    -0.1473218   0.0472494  -3.1180 0.002572 **
panel3$x7     0.2105368   0.1063845   1.9790 0.051437 .
panel3$x8     0.0411521   0.0340840   1.2074 0.231034
panel3$x9    -0.0018695   0.0010833  -1.7258 0.088455 .
panel3$x10   -0.9248687   1.0800808  -0.8563 0.394526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.131
Residual Sum of Squares: 0.091503
R-Squared:                0.3015
Adj. R-Squared: 0.20959
F-statistic: 3.28042 on 10 and 76 DF, p-value: 0.0014222

```

1. Какой пакет позволяет создавать в R данные в виде динамических рядов?
2. Какая функция позволяет сделать декомпозицию ряда динамики?
3. Какие модели строятся по рядам динамики в R?
4. Как выбрать лучшую модель тренда?

1. Эконометрика и эконометрическое моделирование в EXCEL и R: учебник/Л.О. Бабешко, И.В. Орлова. – Москва: ИНФРА-М, 2021. – 300 с.: ил. – (Высшее образование: Магистратура). –DOI 10.12737/1079837.
2. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил. ISBN 978-5-947060-077-1