Линейный регрессионный анализ

Примеры. Прогноз продаж красного вина в Австралии

Продажи вин в Австралии

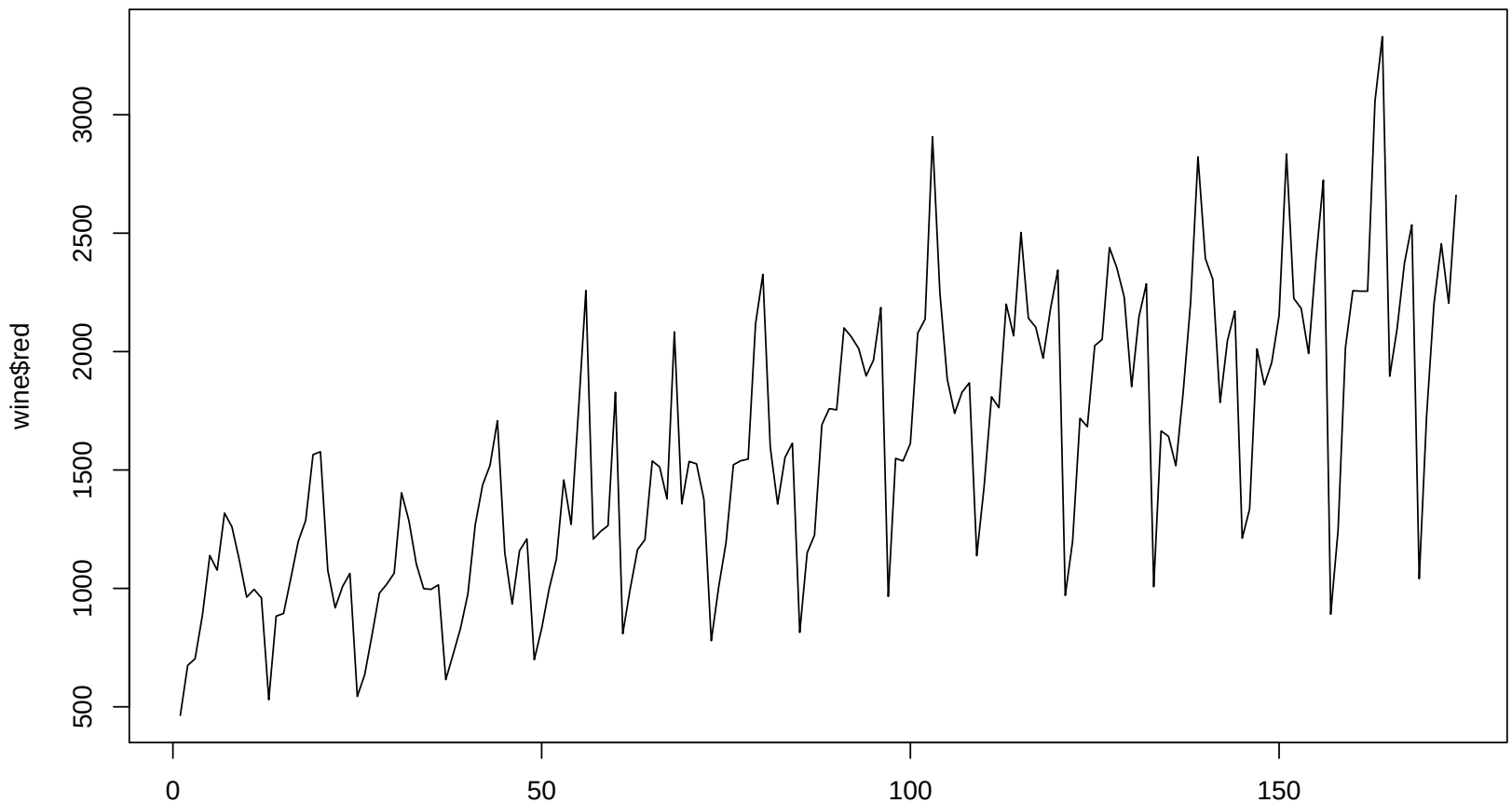
Ежемесячные данные с января 1980 по июнь 1994.

- fort - крепленые (тысячи литров)
- dry - сухие вина (тысячи литров)
- sweet - сладкие вина (тысячи литров)
- red - красные вина (тысячи литров)
- rose - розовые вина (тысячи литров)
- spark - игристые вина (тысячи литров)
- total - общие продажи вин производителями в бутылках объемом не более одного литра

Необходимо построить прогноз на 8 месяцев.

Читаем данные и строим график

```
#  Шаг 0. Прочитаем данные. Внимание: разделитель полей Tab!
setwd("week_08/data/")
wine <- read.table("wine_Austral.dat", header=T, sep="\t")
#  Шаг 1. Предварительный анализ: построим график ряда
plot(wine$red, type="l")
```



Результаты анализа

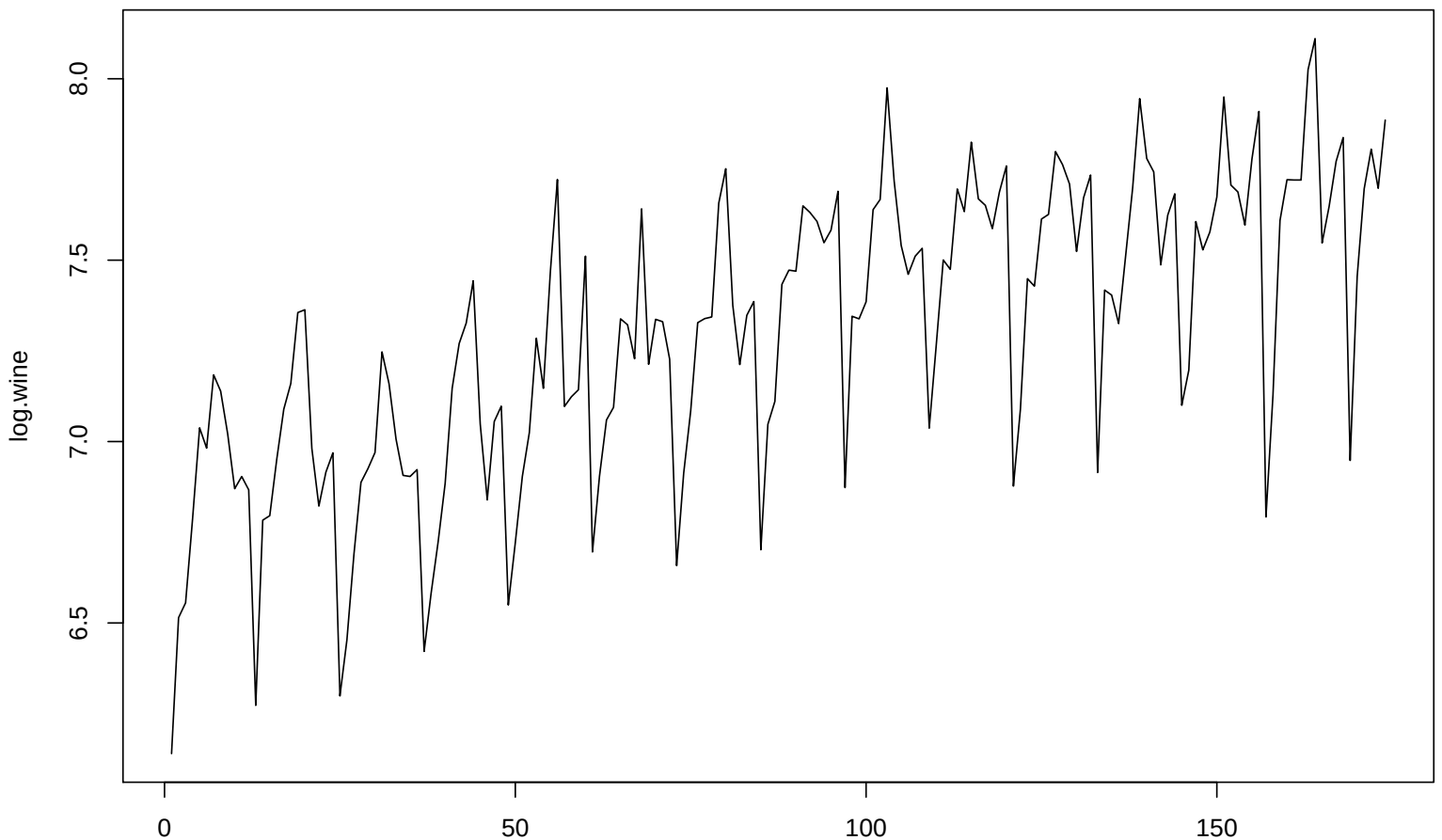
Из графика видно, что:

- Тренд есть, тренд линейный.
- Сезонность есть, мультипликативная.

Чтобы получить аддитивную сезонность, можно попробовать рассмотреть логарифм ряда.

Шаг 2. Преобразование временного ряда

```
log.wine <- log(wine$red)
#  Посмотрим результат на графике
plot(log.wine, type="l")
```



Анализ логарифмированных данных

- Тренд есть, тренд примерно линейный.
- Сезонность есть, аддитивная.

Выводы:

- Преобразование привело к желаемому результату.
- Можно строить регрессионную модель с линейным трендом

Шаг 3. Создание дополнительных переменных

Создаем независимые переменные. Делаем это с запасом на те месяцы, для которых будет строиться прогноз

```
len <- nrow(wine)+8 # Нужно 175+8 строк
#  Время
time <- 1:len
#  Сезонные индикаторы
month <- as.factor(rep_len(1:12, len))
#  Чтобы уравнять длины всех векторов
#  добавим к исходным данным пропущенные значения
log.wine[175:len] <- NA
#  Для удобства работы склеиваем из векторов таблицу
wine.02 <- data.frame(log.wine, time,  month)
```

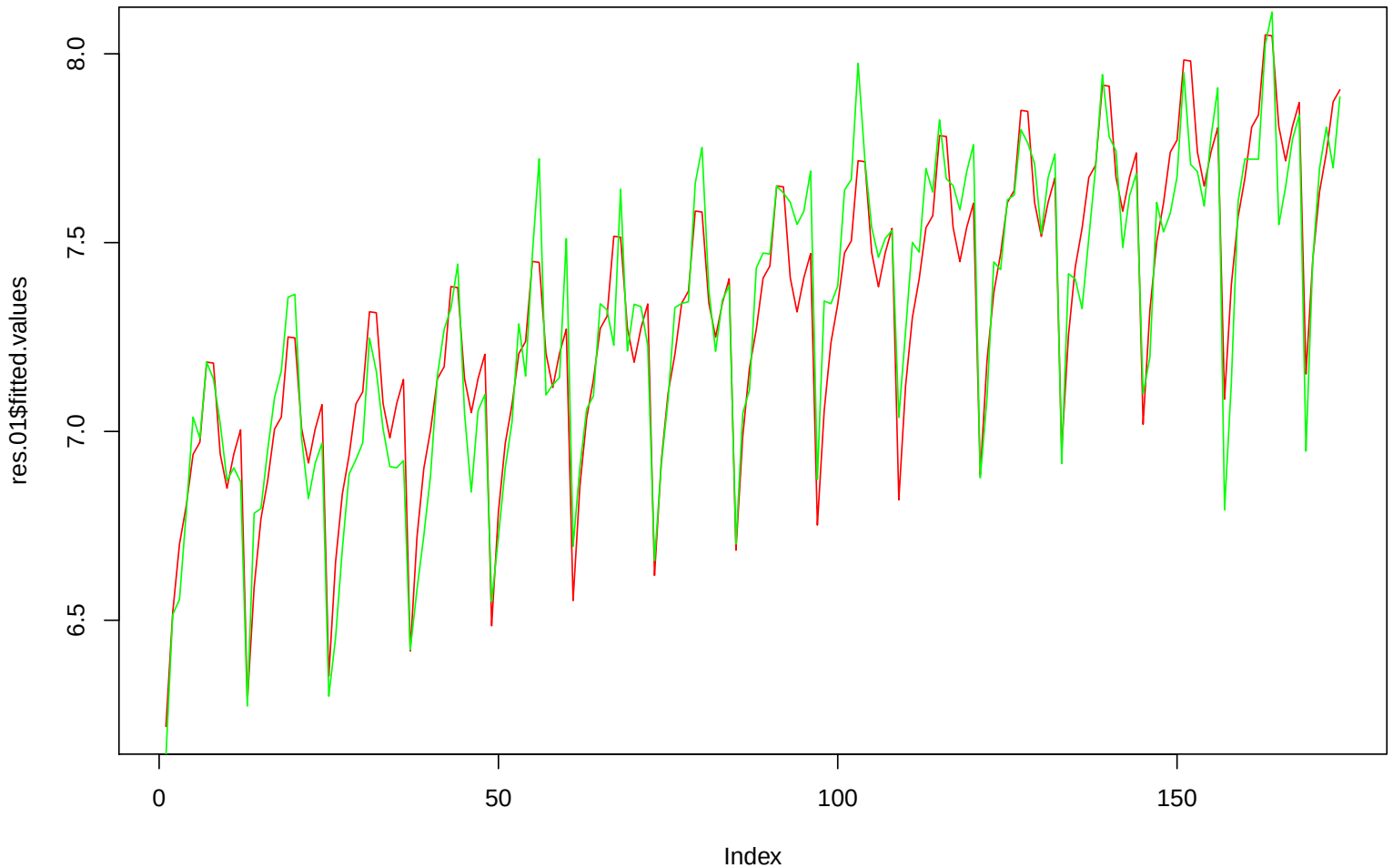
Шаг 4. Регрессионный анализ

```
#  Линейная регрессия. За базу автоматически берется январь
res.01 <- lm(log.wine ~ . , wine.02)
#  Просмотр результатов
summary(res.01)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.2131856  0.0348337  178.37  < 2e-16 ***
time        0.0055557  0.0001824   30.45  < 2e-16 ***
month2      0.2968686  0.0441120    6.73 2.83e-10 ***
month3      0.4719070  0.0441131   10.70  < 2e-16 ***
month4      0.5686400  0.0441150   12.89  < 2e-16 ***
month5      0.6984818  0.0441176   15.83  < 2e-16 ***
month6      0.7248984  0.0441210   16.43  < 2e-16 ***
month7      0.9315212  0.0448924   20.75  < 2e-16 ***
month8      0.9233443  0.0448928   20.57  < 2e-16 ***
month9      0.6769936  0.0448939   15.08  < 2e-16 ***
month10     0.5809970  0.0448957   12.94  < 2e-16 ***
month11     0.6657786  0.0448983   14.83  < 2e-16 ***
month12     0.7244539  0.0449017   16.13  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Шаг 5. Подгонка для логарифма ряда

```
plot(res.01$fitted.values,  type="l", col="red")
lines(wine.02$log.wine, col="green")
```

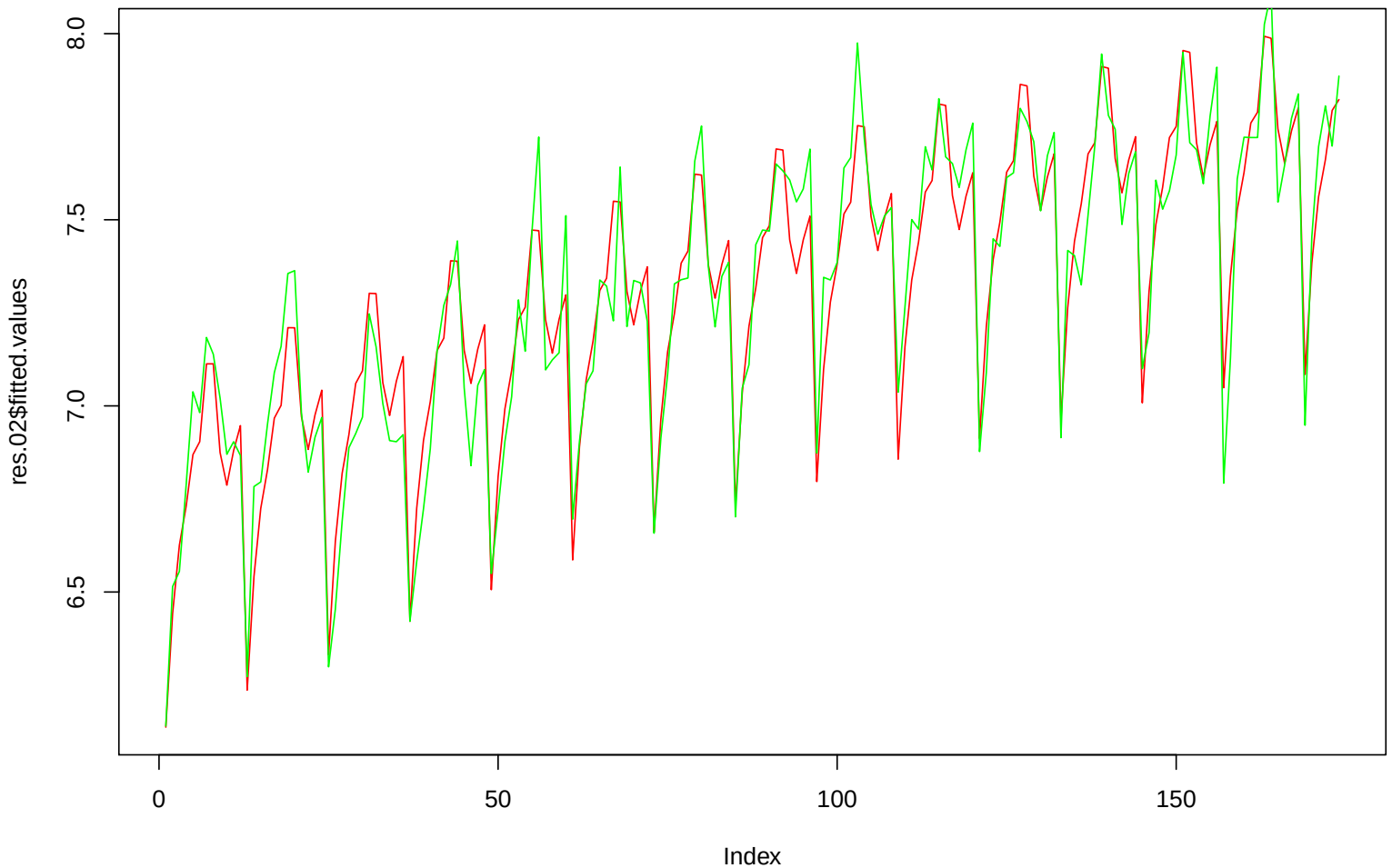


Попытка с параболическим трендом

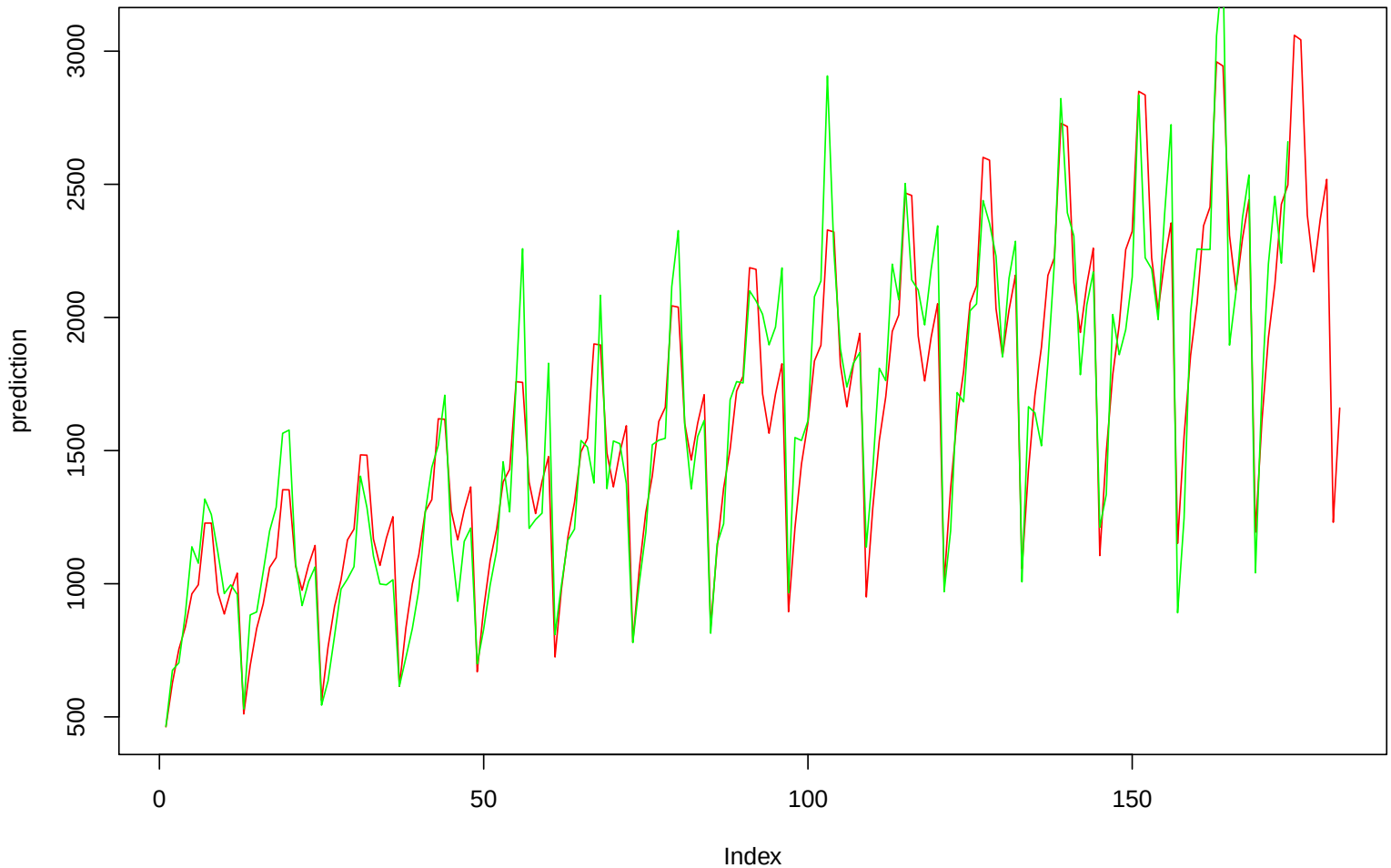
```
#  Время
time <- 1:len
time2 <- time*time
...
wine.03 <- data.frame(log.wine, time, time2, month)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.128e+00  3.811e-02 160.795  < 2e-16 ***
time         8.549e-03  6.960e-04  12.283  < 2e-16 ***
time2       -1.711e-05  3.853e-06  -4.440 1.67e-05 ***
month2       2.968e-01  4.175e-02   7.109 3.67e-11 ***
month3       4.718e-01  4.175e-02  11.300  < 2e-16 ***
month4       5.685e-01  4.176e-02  13.616  < 2e-16 ***
month5       6.984e-01  4.176e-02  16.725  < 2e-16 ***
month6       7.249e-01  4.176e-02  17.358  < 2e-16 ***
month7       9.256e-01  4.251e-02  21.772  < 2e-16 ***
month8       9.173e-01  4.251e-02  21.577  < 2e-16 ***
month9       6.709e-01  4.251e-02  15.781  < 2e-16 ***
month10      5.749e-01  4.252e-02  13.523  < 2e-16 ***
month11      6.598e-01  4.252e-02  15.517  < 2e-16 ***
month12      7.185e-01  4.252e-02  16.898  < 2e-16 ***
```

График с параболическим трендом



Итоги



Возможно следует отрезать первые 4 года наблюдений и это позволит точнее прогнозировать размах сезонных колебаний на оставшейся части ряда.