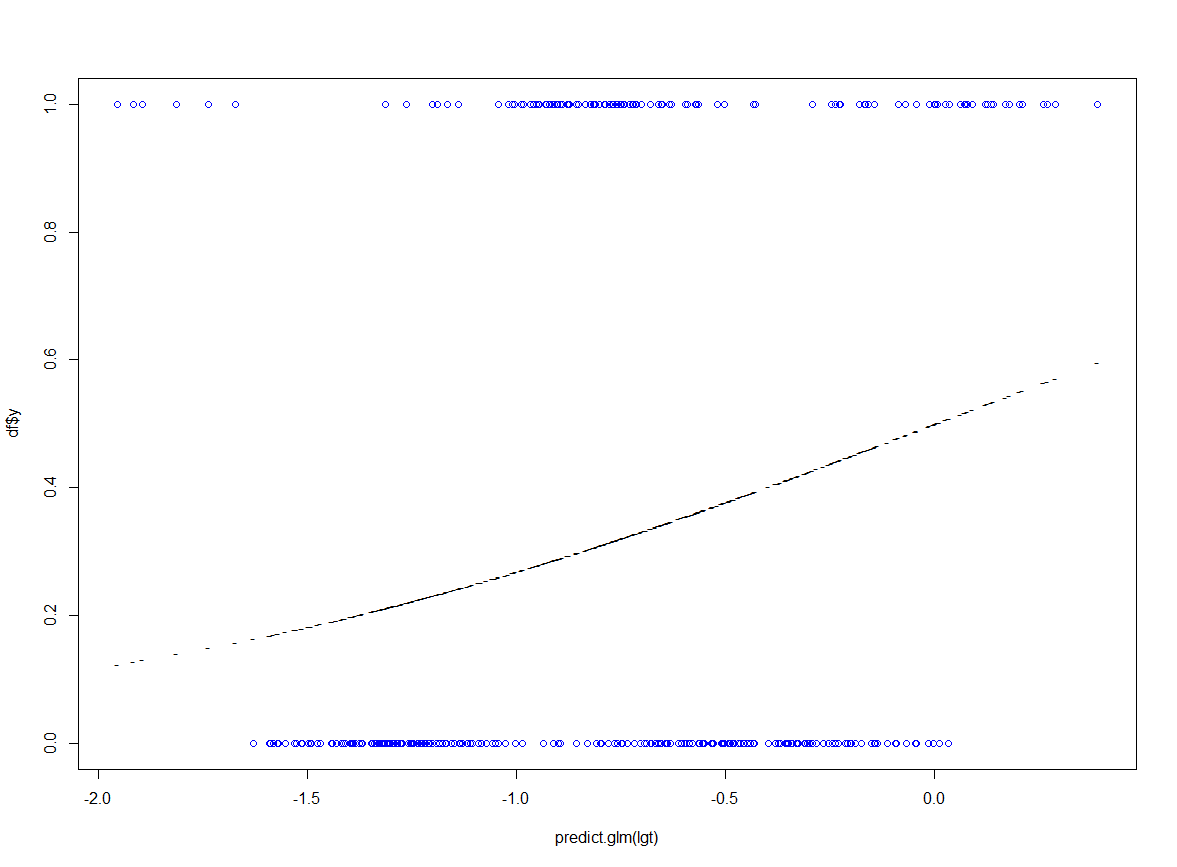
Фролов Максим 21712

**Задание 1**

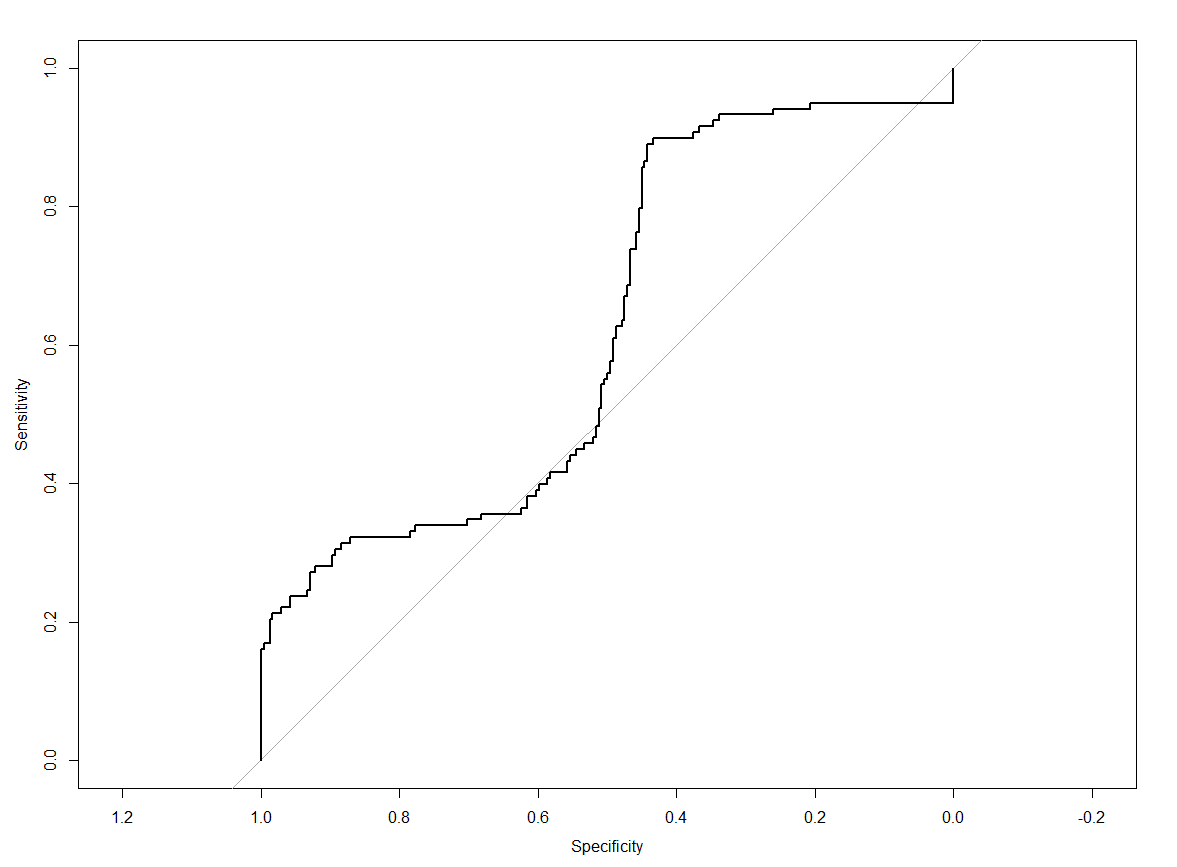
**1а) Загрузите данные по одной объясняемой переменной y (вида 0/1) и двум объясняющим x1, x2. Оцените модель логит. Постройте кривую ROC для имеющихся данных (не прогнозов).**

Построим регрессию логит с помощью функции glm от переменных x1 и x2. lgt=glm(df$y~df$x1+df$x2, family = "binomial"). Ниже приведен график для наблюдений с вероятностями для исходных данных. С помощью summary в скрипте можно посмотреть коэффициенты регрессии. X1 на уровне значимости 0.05 незначима в регрессии, p-value = 0.0504. Константа тоже незначима, p-value = 0.5564.



Ниже приведена ROC кривая для исходных данных, не прогнозов. (plot.roc(df$y, fitted(lgt)))

Вывод: кривая выше, чем прямая линия y=x, но площадь под ROC кривой визуально не сильно больше 0.5 из-за того, что кривая проваливается примерно в середине, прогноз не очень хороший.

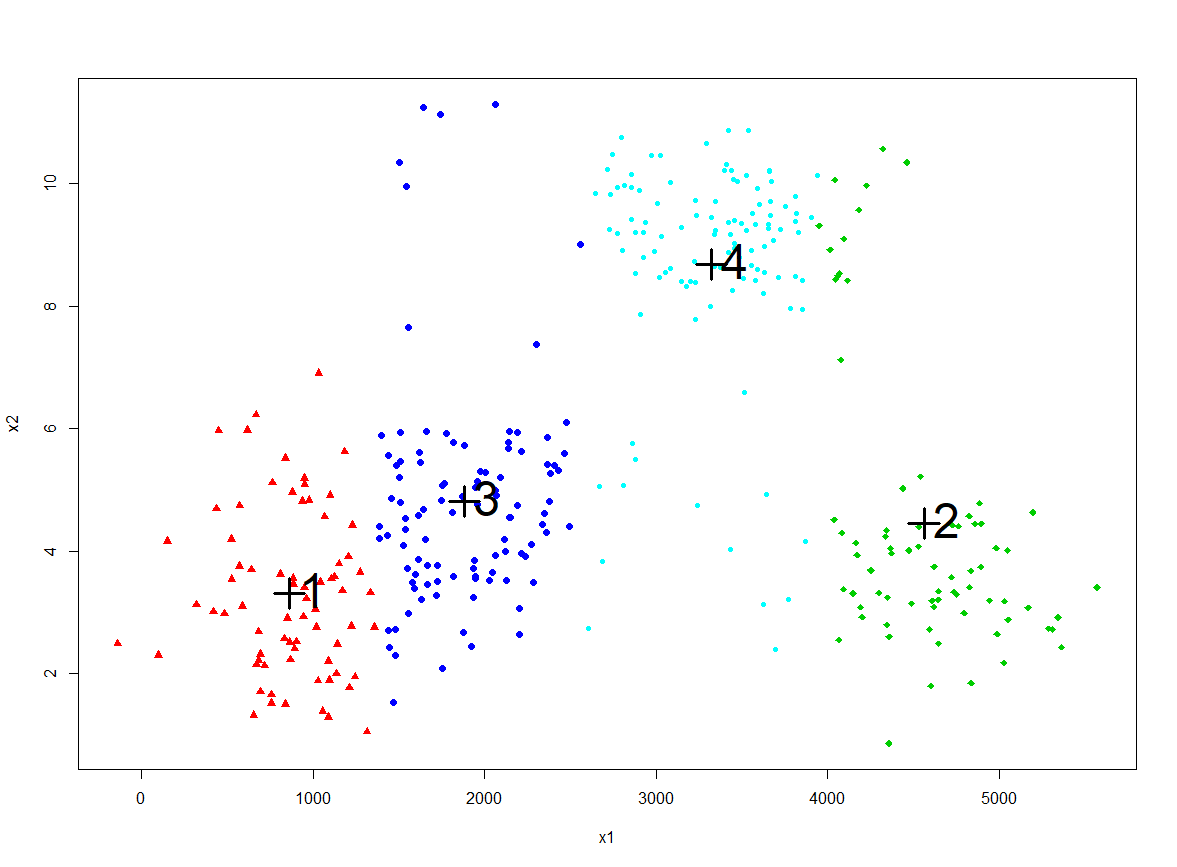


**1б) Разбейте точки x1, x2 на кластеры. Сами выберите число кластеров и параметры. Изобразите кластеры на графике. Создайте категориальную переменную принадлежности кластеру.**

В новом фрейме была удалена переменная y, чтобы только остались x1 и x2. Я выбрал 4 кластера и 50 итераций для kmeans. Ниже приведен график зависимости x2 от x1 и отображены на нем, получившиеся кластеры. Крестиком обозначены центры, также кластеры подписаны и выделены разным цветом. Наблюдаем визуально ровные границы между областями, кластеры друг на друга не заходят.

Была создана факторная переменная z принадлежности кластеру и добавлена в фрейм.

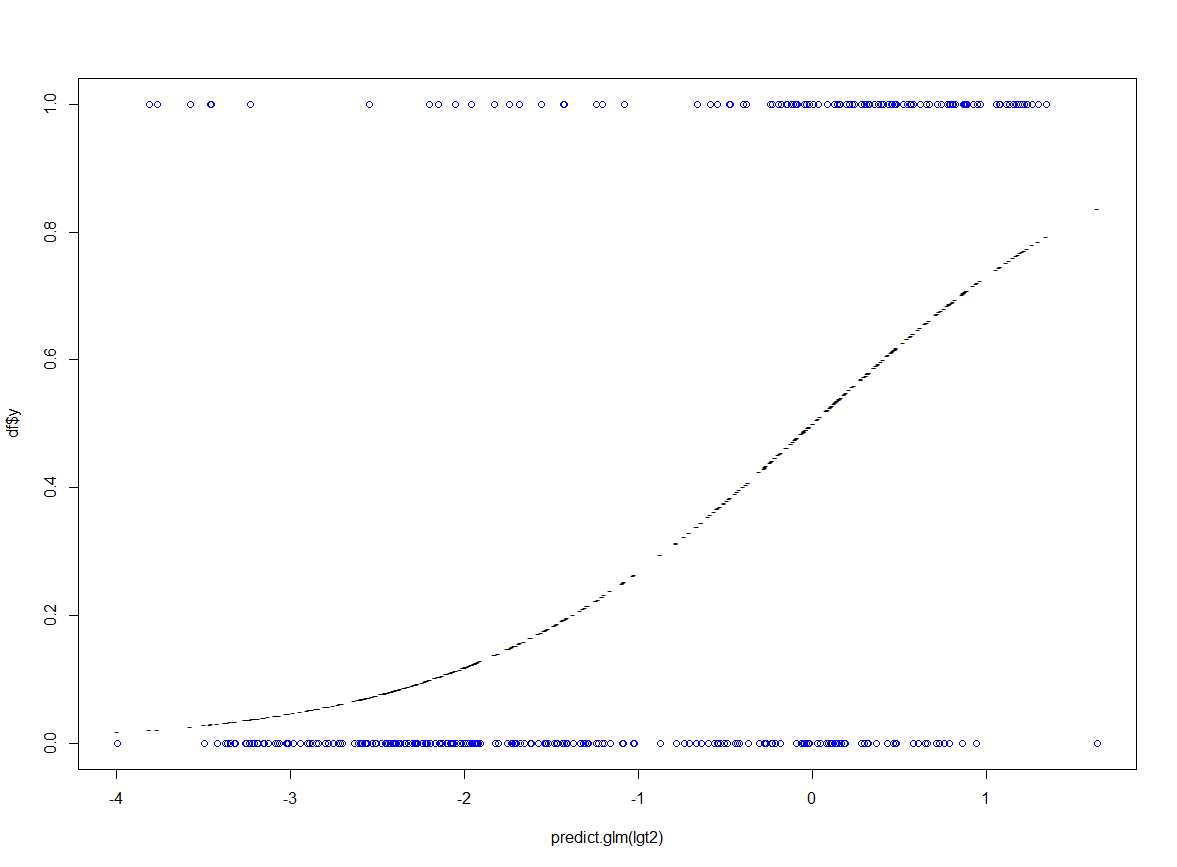
df$z <- factor(km$cluster)



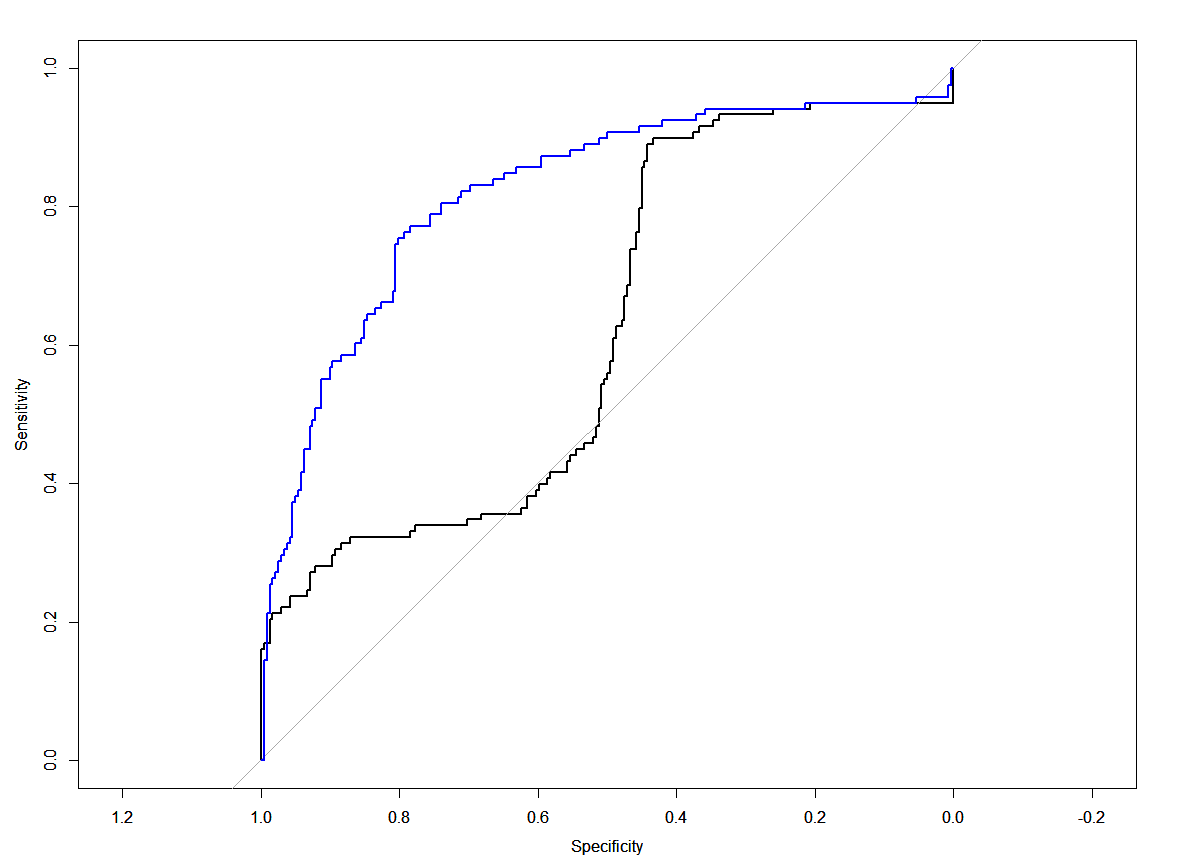
**1в) Оцените модель логит, добавив в 1-ю модель созданную категориальную переменную. Постройте аналогичную кривую ROC и добавьте на тот же график ROC по 1-й модели.**

Была построена новая модель с учетом созданной категориальной переменной z. lgt2=glm(df$y~df$x1+df$x2+df$z, family = "binomial")

Ниже приведен график наблюдений с вероятностями для новой модели.



Была построена ROC кривая для 2 модели. На графике ниже синим цветом обозначена кривая для 2 модели, черным для 1. Вывод: по метрике ROC визуально видно, что 2 модель точнее, площадь под ROC кривой визуально больше и сама кривая значительно отходит от прямой линии.



**1г) По 2-й модели запишите формулу, взяв конкретные коэффициенты из таблицы, по которой можно рассчитать прогноз вероятности события y=1 для некоторых x1, x2 и 2-го по счету кластера.**

Ln() = -7,68+0,00135x1+0,0627x2+5,3064z2+4,65274z3+1,16408z4.

Для x2 p-value=0.37, x2 назначим при 5% уровне значимости. Z4 p-value 0.07, тоже назначим для 5%.

**1д) Постройте по 2-й модели таблицу сопряжённости (contingency она же confusion) для уровня вероятности 0.5.**

table(df$y, fitted(lgt2) > 0.5)

|  |  |  |
| --- | --- | --- |
|  | False | True |
| 0 | 206 | 36 |
| 1 | 44 | 74 |

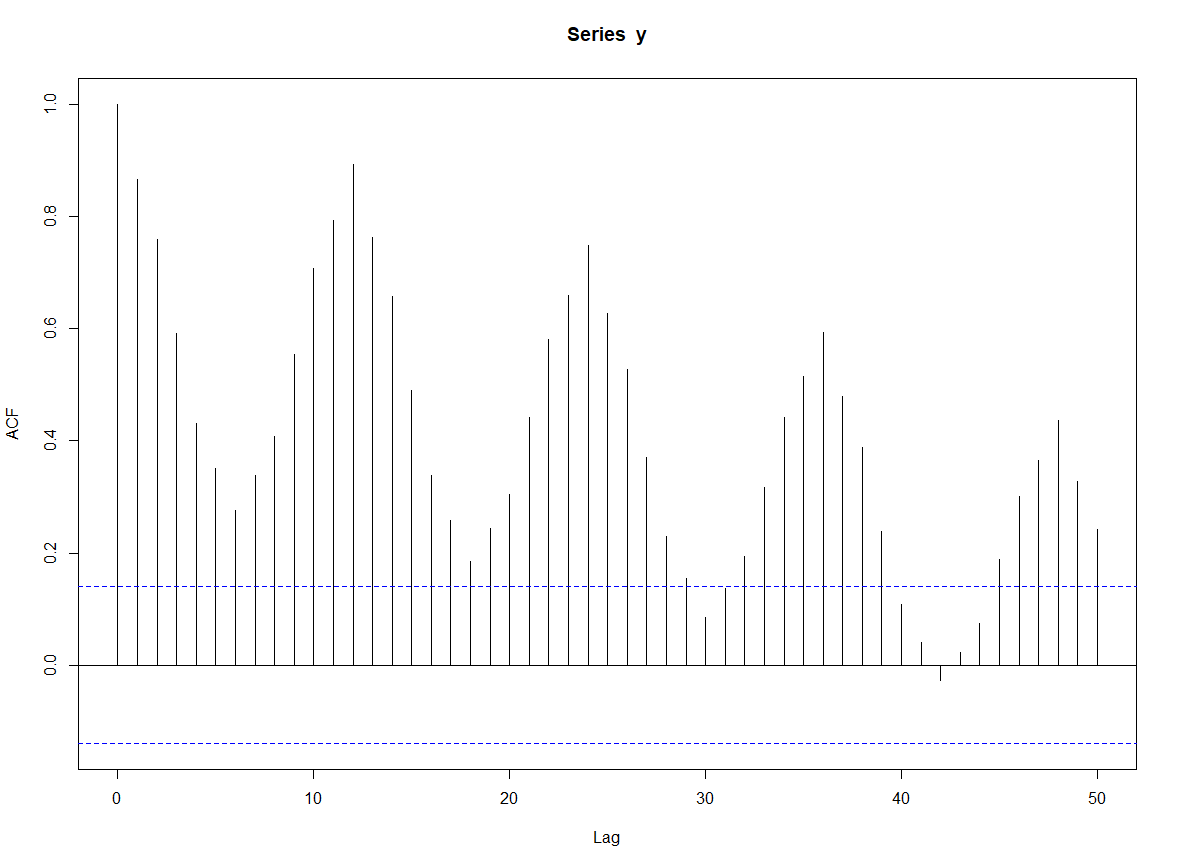
**1е) Как вы думаете, какая модель более точно описывает данные? Объясните.**

Если судить по метрике ROC, то вторая модель (это и логично, потому что она учитывает больше факторов, например факторную переменную z, когда мы били x1 и x2 на 4 кластера, а это больше, чем 2). Площадь под ROC кривой визуально больше для 2 модели и сама кривая значительно отходит от прямой линии.

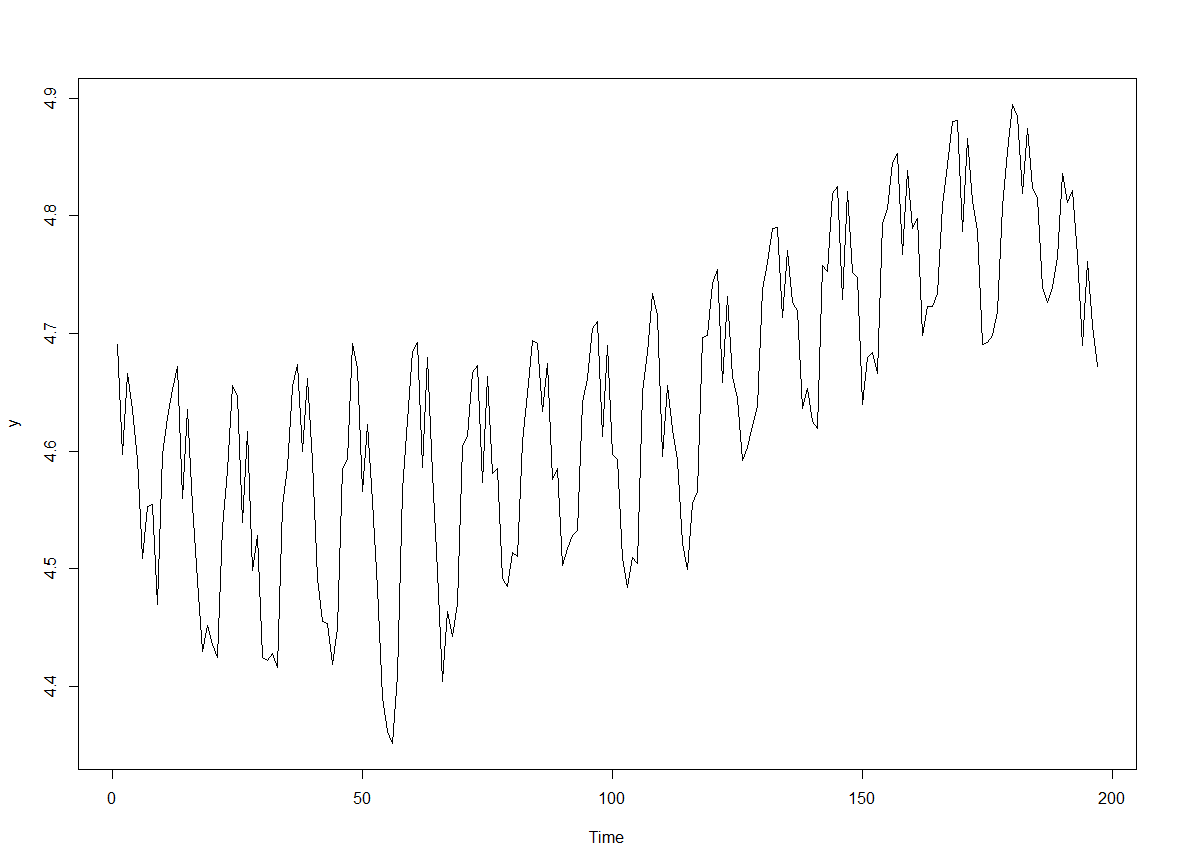
**Задача 2**

**2а) Постройте каким-либо методом прогноз ряда, взяв последние 24 наблюдения в качестве тестового периода, а все предыдущие в качестве тренировочного периода. Прогноз должен быть осмысленным, но не обязательно точным. Объясните свой метод словами.**

Я взял сезонные данные GAZ\_C и для них построил прогноз. Сначала я логарифмировал значения и решил построить АКФ, она приведена ниже. Видно, что имеется сезонность на АКФ.

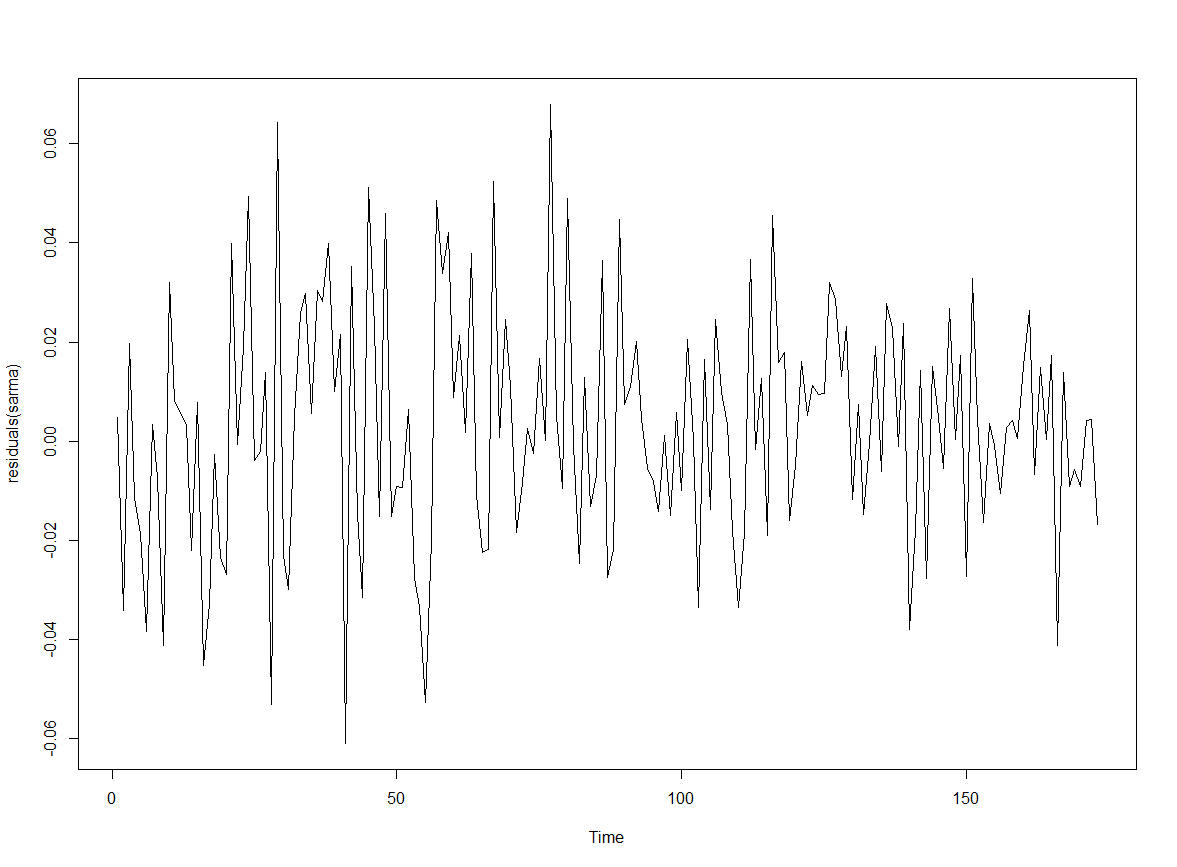


Ниже приведен график самих значений данных, видим сезонность.



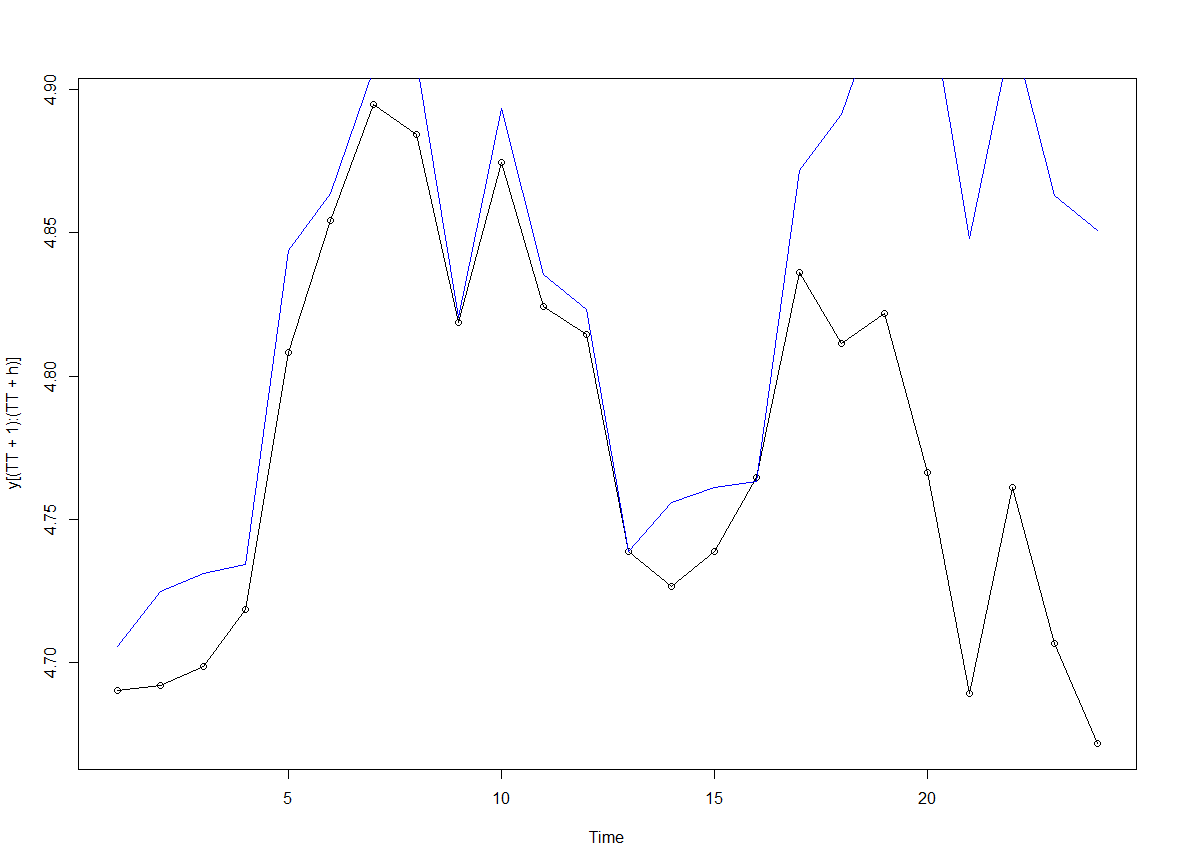
Использовал модель arima, в качестве тренировочного периода взял все, кроме последних 24 месяцев. arima(y[1:TT], order=c(1,1,1), seasonal = list(order = c(1, 0, 1), period=12)). Порядок arima – p=1, d=1, q=1. d=1 – взяли одну разность для стационарности, и используем по одному лагу для AR и MA. Сезонность у нас не меняется – стационарная.

Ниже приведен график белого шума, который мы оценили – остатки.

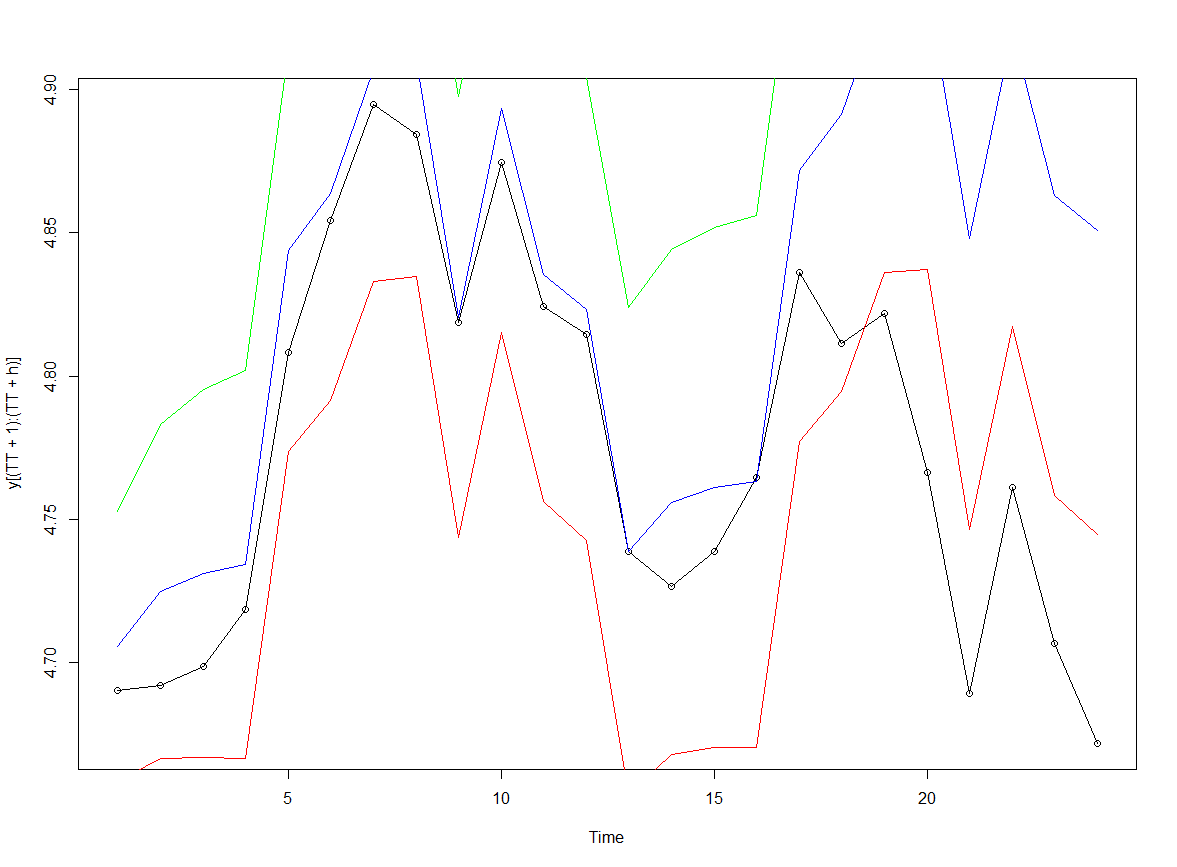


**2б) Постройте график фактического значения ряда и прогноза. Рассчитайте ошибки прогноза и нарисуйте их график (+ график нуля).**

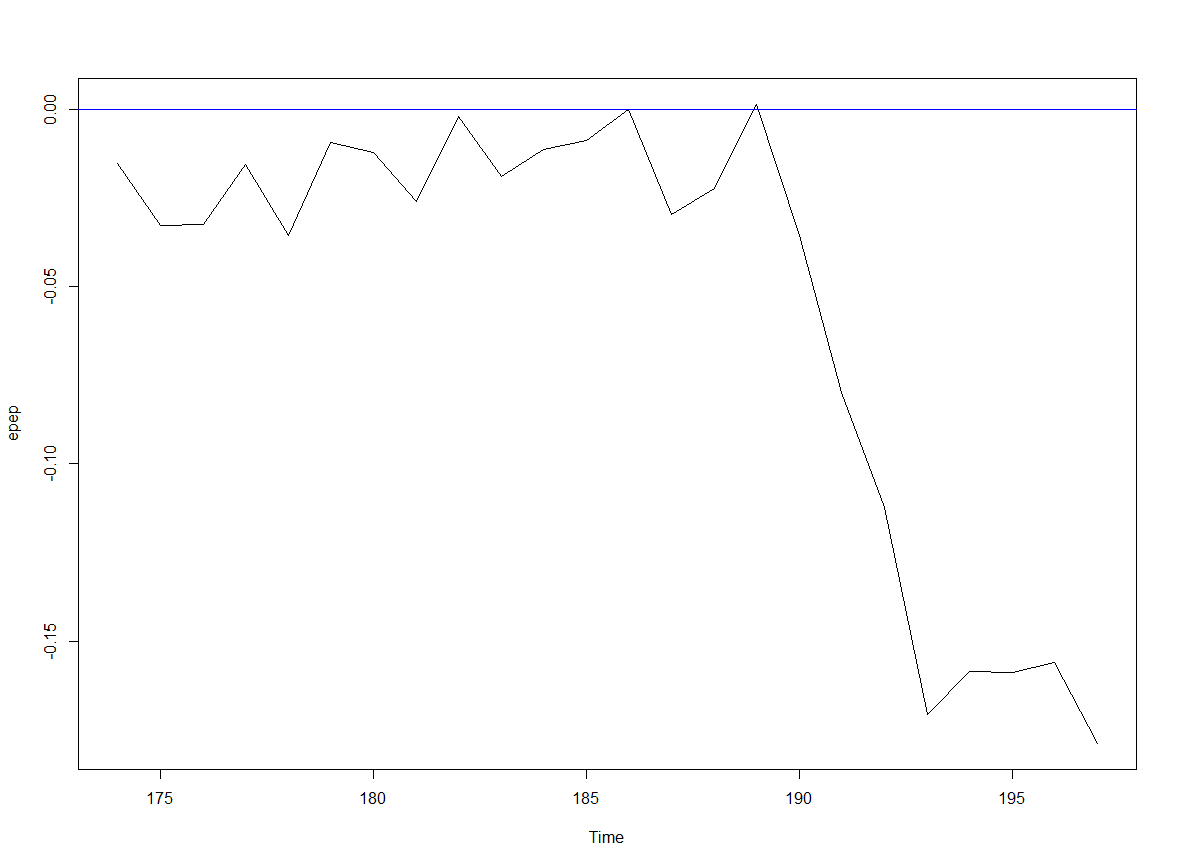
Ниже построен график фактического значения тестового ряда – последние 24 месяца (значения) и прогноза. Прогноз – синим цветом. Черным – исходные данные.



Был построен 5% интервальный прогноз. Зеленый- верхняя граница, красный – нижняя. Визуально видно, что в самом конце черная линия выходит за доверительный прогноз.



Были вычислены ошибки (остатки) и нарисован график, который приведен ниже. Линия нуля – синим цветом. Ошибка считались как из данных вычитал прогнозные значения. Визуально видим, что точечный прогноз оказался завышенным, особенно в конце периода.



**2в) Рассчитайте показатель MAD для вашего прогноза.**

Были посчитаны два показателя, MAE и RMSE.

MSE <- mean(epep^2), RMSE <- sqrt(MSE) = 0.082378

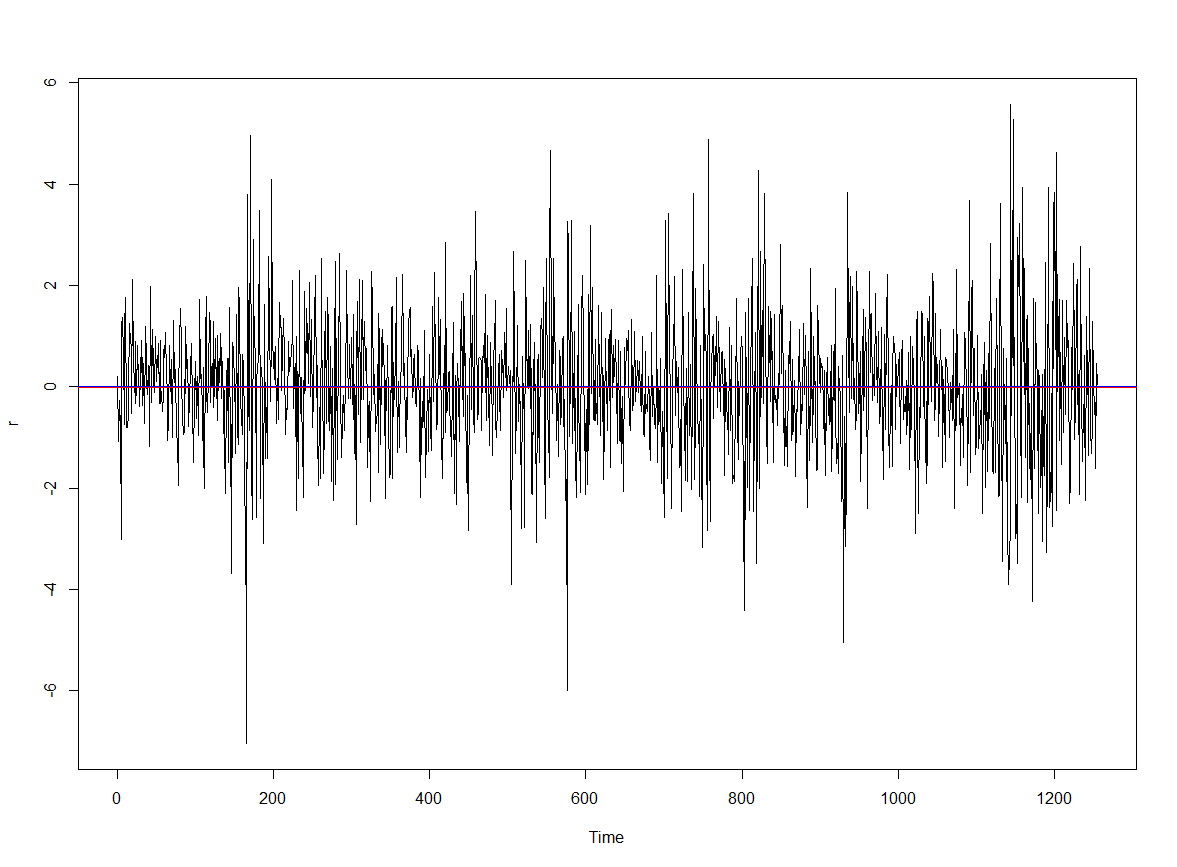
MAE <- mean(abs(epep)) = 0.0551917

**Задача 3**

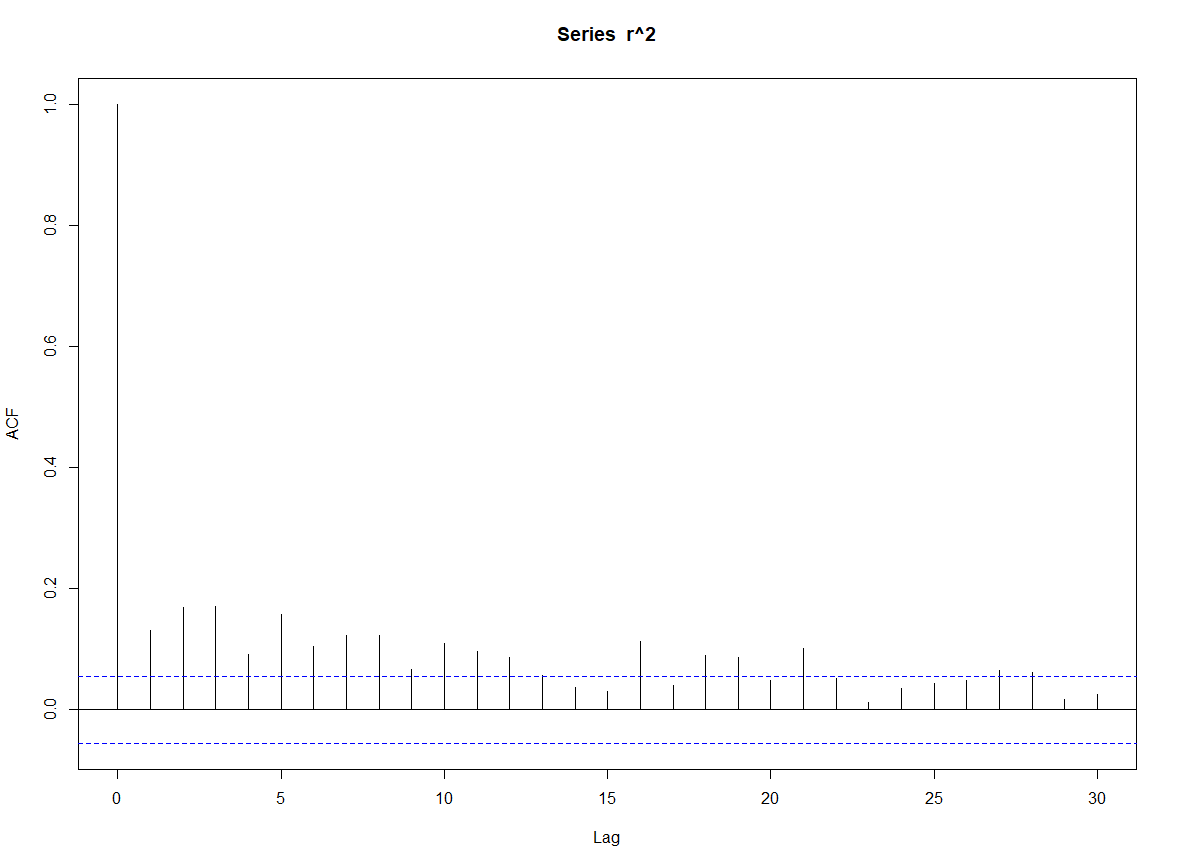
**3а) Возьмите свой ряд значений фондового индекса и создайте ряд доходностей. Постройте график ряда и убедитесь каким-либо способом, что имеет место кластеризация волатильности.**

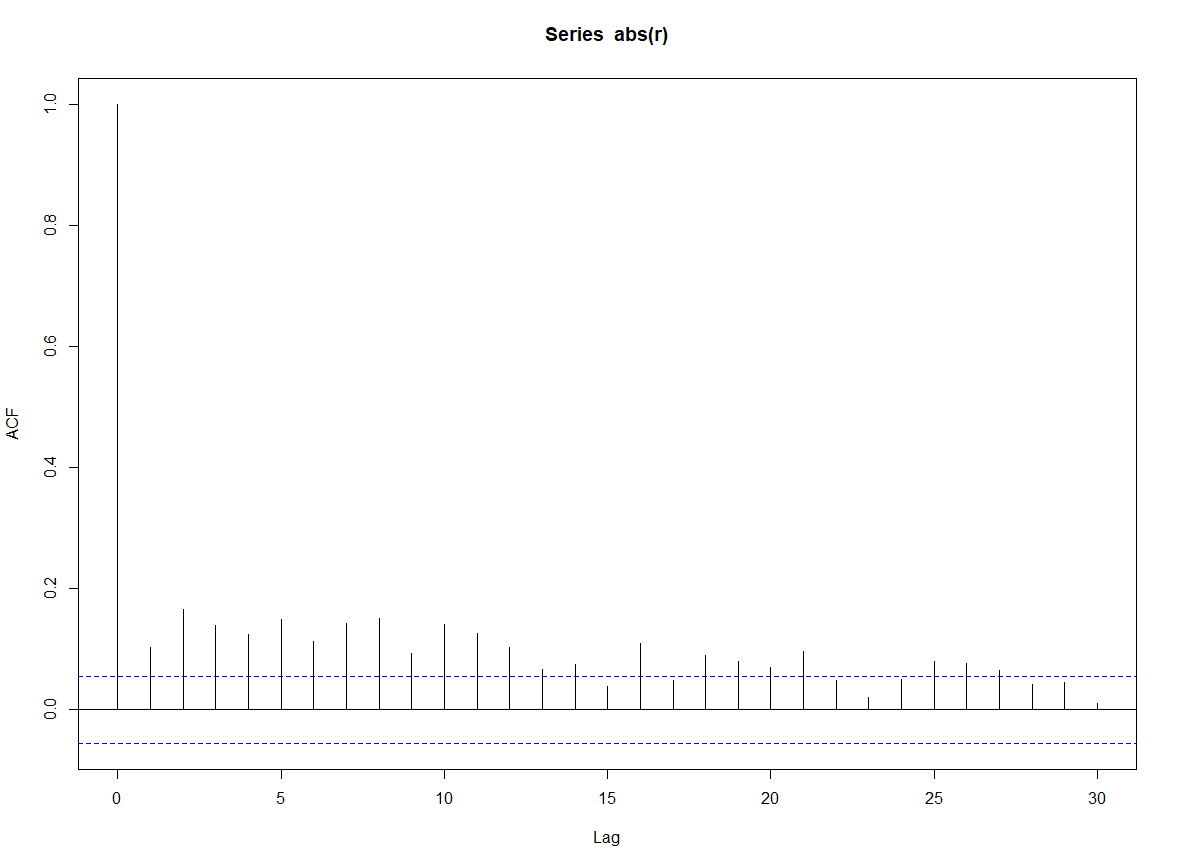
Мой ряд был с 1998 до 2002 включительно. Считал логарифмические доходности в процентах. r <- diff(log(df1$Close)) \* 100

Ниже приведен график логарифмической доходности в % за весь период. Синий – линия 0, красный – среднее, визуально видим, что в среднем колеблется вблизи 0.



Был построен АКФ для r^2 и модуля r. Графики приведены ниже. Видим визуально, что имеется кластеризации волатильности. АКФ выше доверительного интервала для белого шума (синий пунктир).



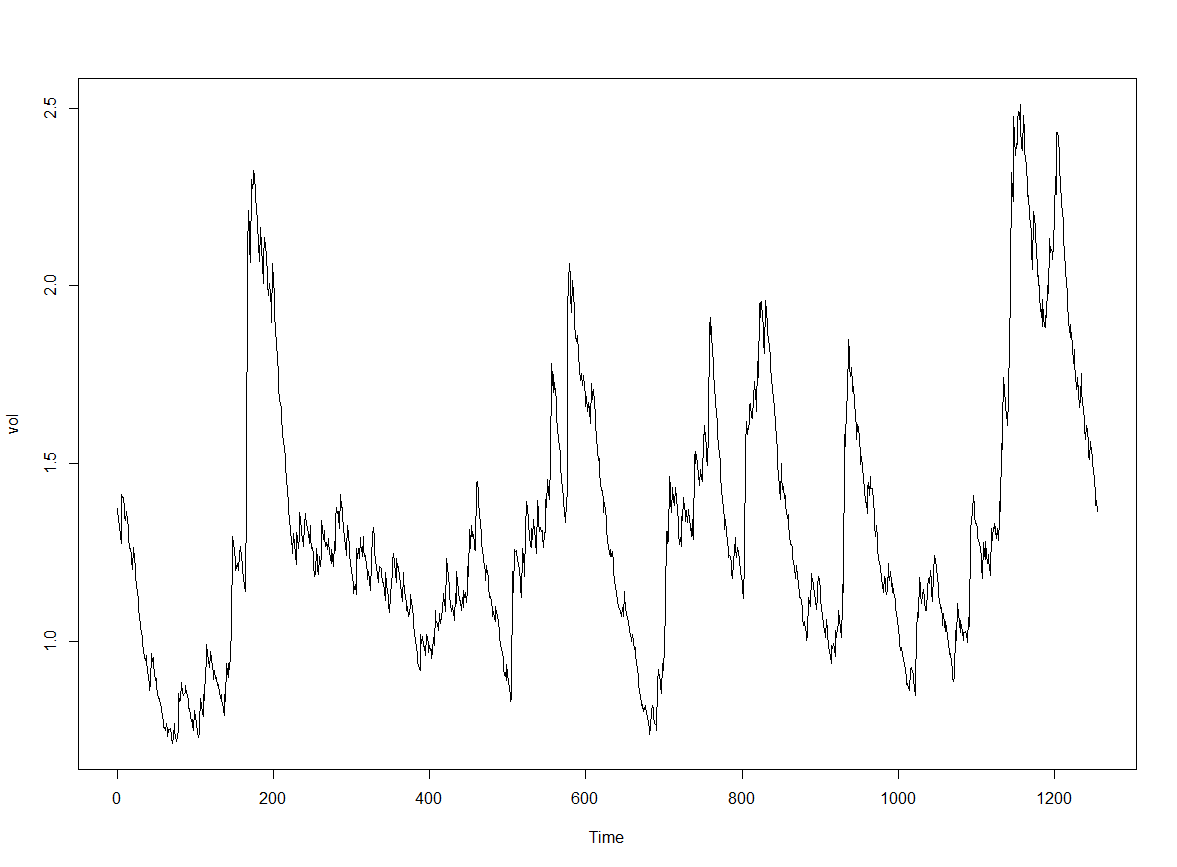


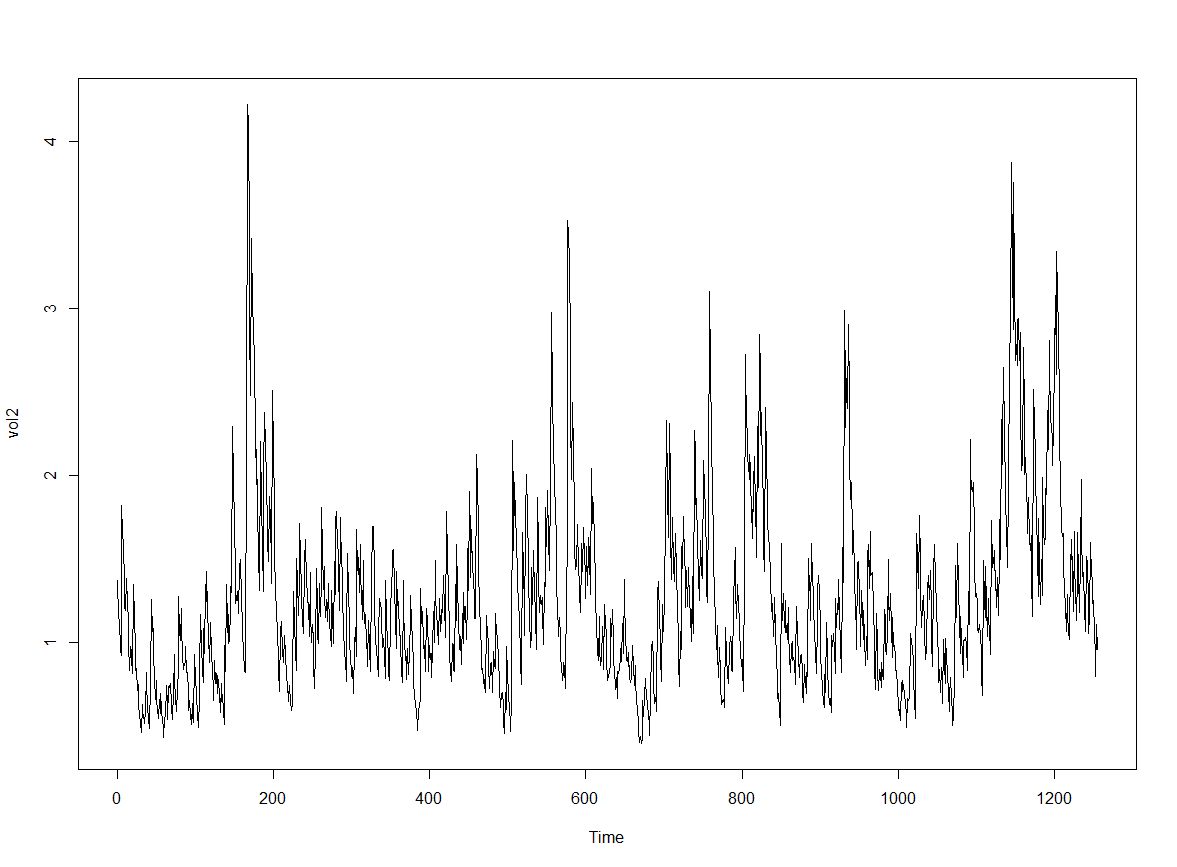
**3б) Оцените волатильность доходностей по экспоненциальному сглаживанию Riskmetrics с какими-либо двумя разными параметрами. Постройте график двух рядов волатильности.**

s1[t] <- a \* r[t-1]^2 + (1-a)\*s1[t-1]

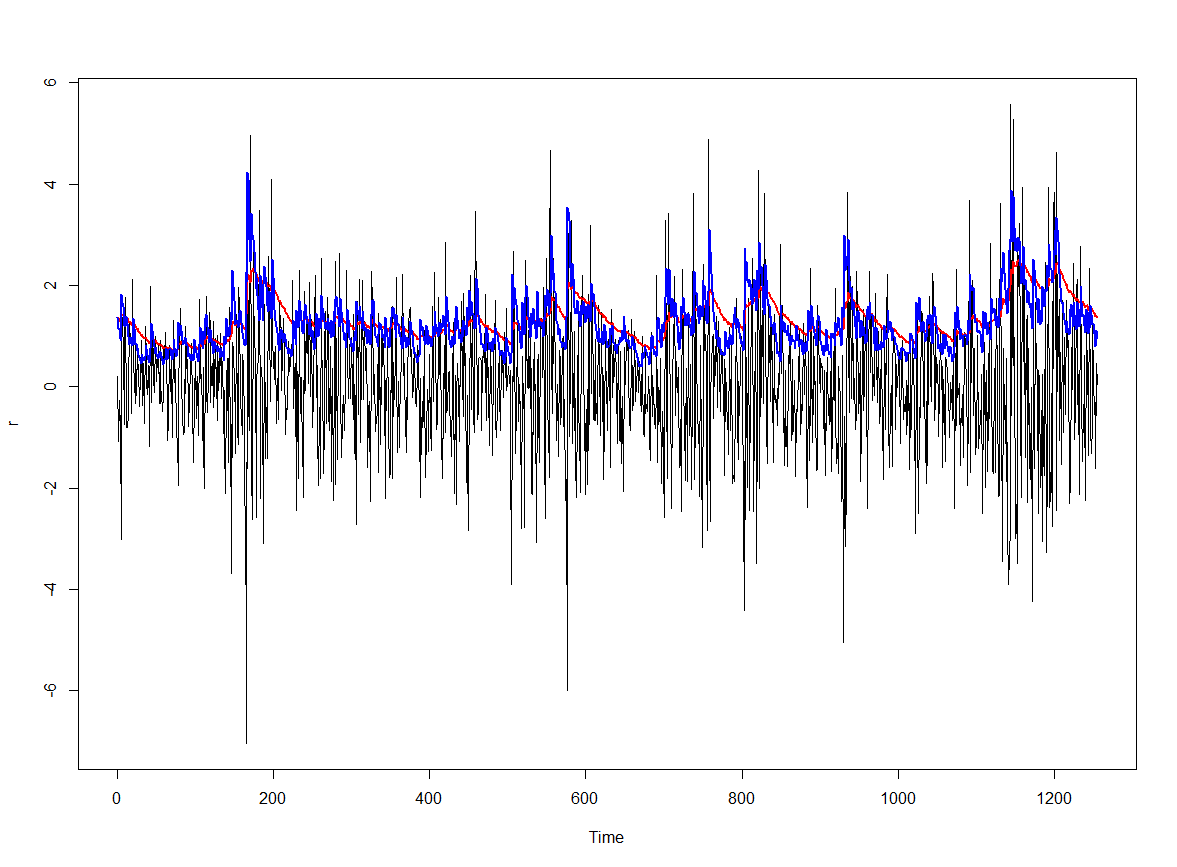
Параметр a был взят для 1 случая – 0.05, для второго 0.3.

На графике ниже приведены волатильности в процентных пунктах, 1 график a = 0.05. Видно, что график с большим параметром a имеет более частый разброс и больше по величине разброс.





На графике ниже изображены черным – доходность, красным – волатильность (a=0.05), синим – волатильность (a=0.3) в процентных пунктах.



**3в) Две оценки волатильности можно рассматривать как прогнозы корня из дисперсии.Получите для двух прогнозов значения потерь по функции потерь**

L1[t] <- (r[t]^2)/(s1[t]) + log(s1[t])

Были посчитаны функции потерь L1(a=0.05) и L2 (a=0.3) для двух оценок волатильности.

Далее посчитал средние потери – среднее по функции потерь.

Средние потери для (a=0.05) = 1.55

Для a=0.3 = 1.69.

То есть у нас получилось, что чем больше a – тем больше потери и менее точная оценка.

library(lmtest)

library(sandwich) подключили библиотеки

Прогнозы были сравнены по тесту Диболда–Мариано. Я построил регрессию разности функций потерь от константы. Используя HAC сделал тест: lmtest::coeftest(reg11, vcov=vcovHAC). P-value 0.000167. То есть волатильность с меньшим a значимо лучше, чем с большим, по крайней мере на уровне значимости 0.001.