Домашнее задание №3

Агалян Роберт БПТ214

Введение

В нашем мини-исследовании будем использовать датафрейм содержащий результаты психолингвистического исследования, посвященного узнаваемости слов, в такого рода исследованиях участникам эксперимента предлагают определить, является ли слово, которое они видят на экране, реально существующим в языке или нет. Мы попытаемся проверить влияет ли длина слова, богатство морфологической семьи слова (как много однокоренных слов с разной частью речи), часть речи и количество синонимов у искомого слова на время, затраченное на узнавание слова (с момента появления слова на экране до нажатия кнопки, реальное слово или нет). В ходе работы попытемся проверить следующие гипотезы: 1. Чем короче слово, тем легче узнать, реальное оно или нет. 2. Чем богаче морфологическая семья слова, тем легче узнать, реальное оно или нет. 3. Определить реальность слова-существительного, сложнее чем глагола. 4. Чем больше количество синонимов, тем легче узнать, реальное оно или нет.

Построение и запуск модели

Сохраним наш датафрейм в **df**, построим и сохраним модель в переменную **fit1**. Зависимой переменной в модели будет , а независимыми переменными будут **LengthInLetters** (длина слова в буквах), **FamilySize** (количество однокоренных слов) + **WordCategory** (часть речи), **NumberSimplexSynsets** (количество слов-синонимов).

```
df <- read.csv("english.csv")
fit1 <- lm(RTlexdec ~ LengthInLetters + FamilySize + WordCategory + NumberSimplexSynse
ts, data = df)
summary(fit1)
```

```
##
## Call:
## lm(formula = RTlexdec ~ LengthInLetters + FamilySize + WordCategory +
##     NumberSimplexSynsets, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -268.92  -82.26   -8.99   69.91  547.66
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           811.030      9.612  84.377   <2e-16 ***
## LengthInLetters         1.995      1.901   1.049    0.294
## FamilySize            -35.042      2.450 -14.304   <2e-16 ***
## WordCategoryV          -1.348      3.449  -0.391    0.696
## NumberSimplexSynsets  -27.665      3.071  -9.010   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106.7 on 4563 degrees of freedom
## Multiple R-squared:  0.1379, Adjusted R-squared:  0.1372
## F-statistic: 182.5 on 4 and 4563 DF,  p-value: < 2.2e-16
```

Проинтепретируем коэффициенты. *Intercept.* Если все независимые переменные равны нулю, и слово - существительное, то среднее время, затраченное на узнавание слова составит 811.030 мс, но ситуация неправдоподобна, потому что в слове не может не быть ни одной буквы.
LengthInLetters. При прочих равных, то есть если мы будем сравнивать время узнавания слов, которые отличаются только длиной, то на слова, которые длиннее на одну букву, будет потрачено в среднем на 1.995 мс больше времени.
FamilySize. При прочих равных, то есть если мы будем сравнивать время узнавания слов, которые отличаются только богатством морфологической семьи (количеством однокоренных слов), то слова, у которых на одно однокоренное слово больше, будут узнаваться в среднем на 35.042 мс быстрее.
WordCategoryV. При прочих равных, то есть если мы будем сравнивать время узнавания слов, отличающихся только частью речи, время, затраченное на узнавание глагола будет в среднем на 1.348 мс меньше, чем на существительное.
NumberSimplexSynsets. При прочих равных, то есть если мы будем сравнивать время узнавания слов, отличающихся только количеством синонимов, то слова, у которых на 1 синоним больше, будут узнаны в среднем на 27.665 мс быстрее.

Если посмотреть на p-value в выдаче, то можем сказать, что оценки коэффициентов при переменной **FamilySize** и **NumberSimplexSynsets** статистически значимы на 0.1% уровне значимости. Остальные оценки оказались незначимыми. Также у нашей модели небольшая предсказательная сила: коэффициент детерминации $R^2 = 13,79\%$. Следовательно, наша модель объясняет 13,79% изменчивости зависимой переменной.

Проверка условий Гаусса-Маркова

Условие о равенстве математического ожидания остатков нулю.

Для начала добавим в датафрейм столбцы с предсказанными значениями и остатками модели и загрузим библиотеку **ggplot2**, она пригодится нам для визуализации:

```
df$res <- fit1$residuals
df$fitted <- fit1$fitted.values
library(ggplot2)
```

Теперь выведем описательные статистики для остатков:

```
summary(df$res)
```

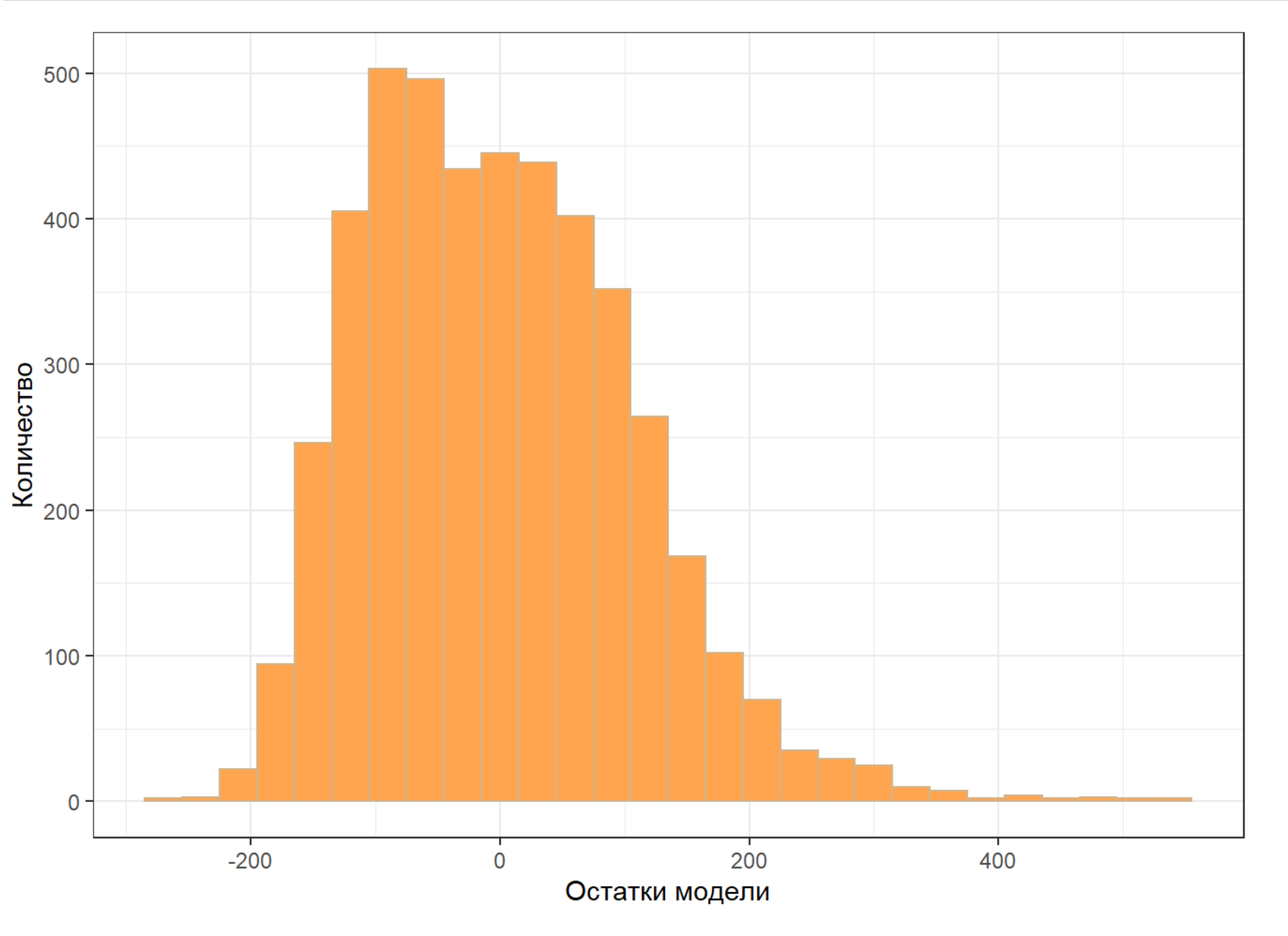
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -268.923  -82.257   -8.987    0.000   69.911  547.655
```

Как мы видим, среднее значени равно нулю, однако медианное совсем нет. Условие не выполняется, попытаемся убедиться в этом далее.

Нормальность распределения остатков

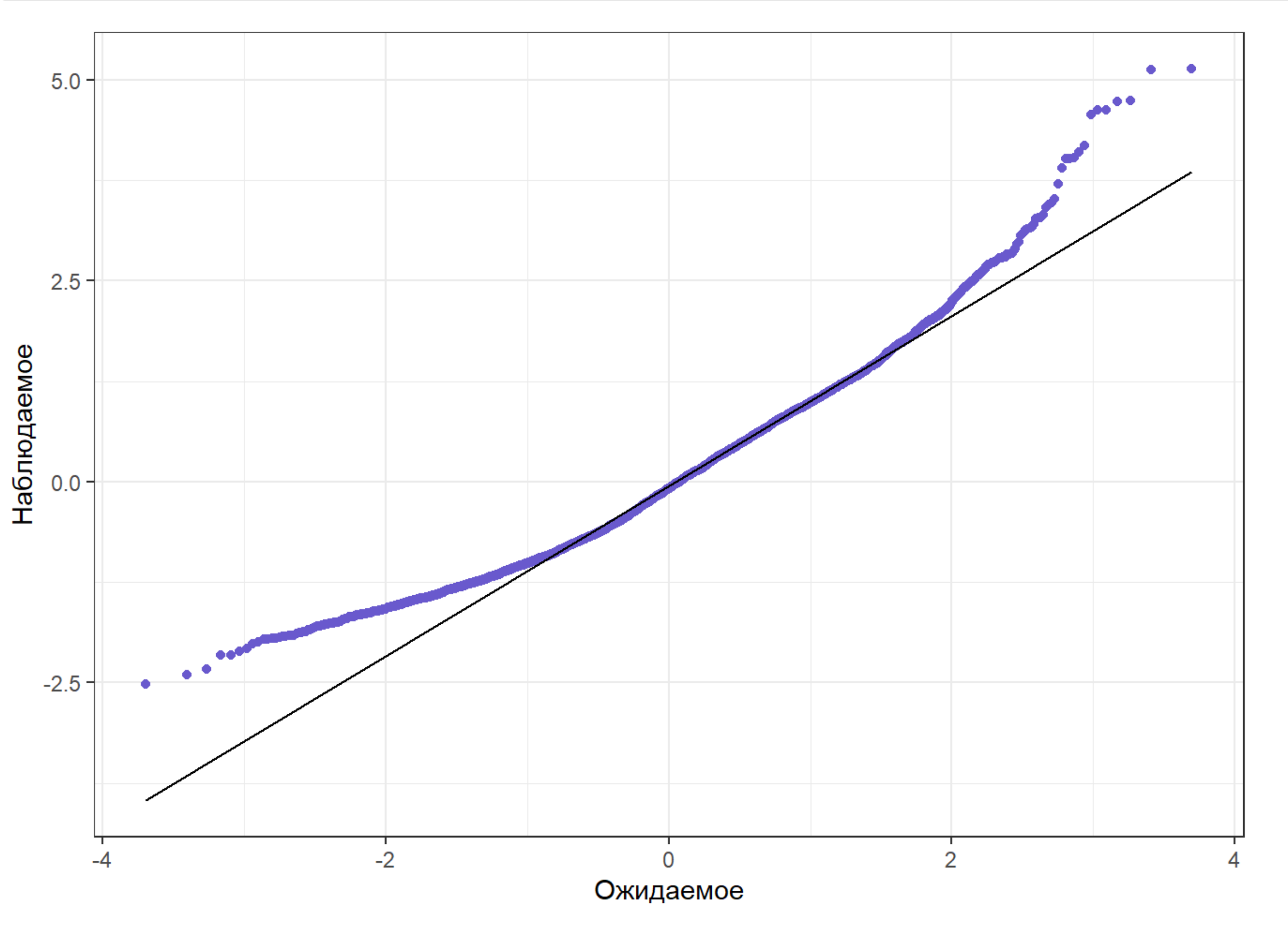
Построим гистограмму по остаткам:

```
ggplot(data = df, aes(x = res))+
  geom_histogram(fill = "tan1", color = "tan", binwidth = 30)+
  xlab("Остатки модели")+
  ylab("Количество")+
  theme_bw()
```



Для лучшего понимания ситуации построим Q-Q plot:

```
ggplot(data = df, aes(sample = scale(res)))+
  stat_qq(color = "slateblue3")+
  stat_qq_line()+
  xlab("Ожидаемое")+
  ylab("Наблюдаемое")+
  theme_bw()
```



Судя по тому, что точки в начале и в конце графика отклоняются от линии, распределение не соответствует нормальному, примени критерий Шапиро-Уилка к нашим остаткам:

```
shapiro.test(fit1$residuals)
```

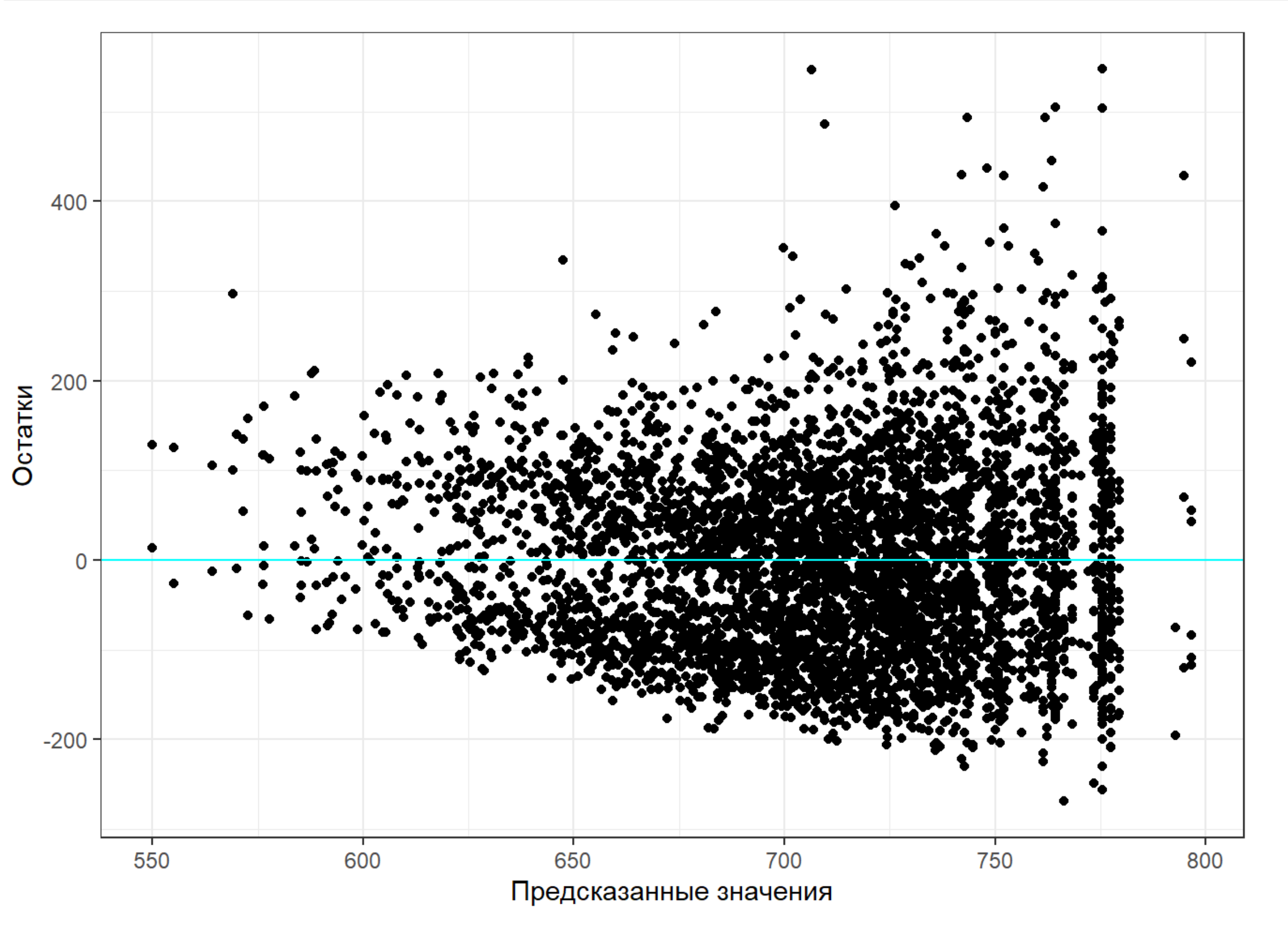
```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.97171, p-value < 2.2e-16
```

P-value мало, отклоняем нулевую гипотезу о нормальности распределения остатков. Условие не выполняется.

Гомоскедастичность

Построим диаграмму рассеивания по предсказанным значениям модели и остаткам:

```
ggplot(data = df, aes(x = fitted, y = res))+
  geom_point() + geom_hline(yintercept = 0, color = "cyan1")+
  xlab("Предсказанные значения")+
  ylab("Остатки")+
  theme_bw()
```



Ошибки остатков заметно больше в правой части диаграммы. Вероятна гетероскедастичность. Загрузим библиотеку **lmtest** для применения критейрия Бройша-Пагана к нашей модели: ##

```
library(lmtest)
bptest(fit1)
```

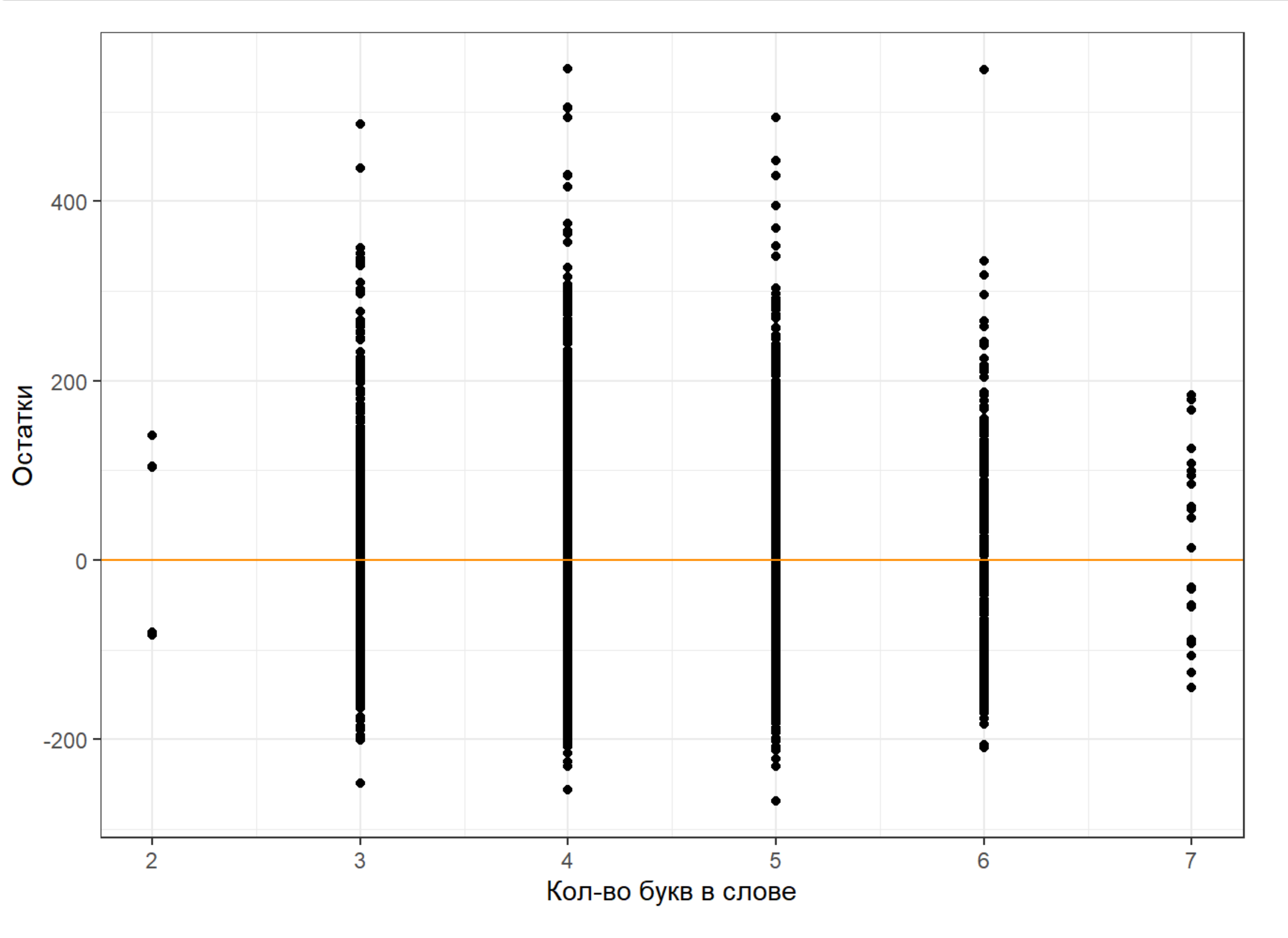
```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 116.04, df = 4, p-value < 2.2e-16
```

P-value мало, отклоняем нулевую гипотезу о гомоскедастичности в пользу альтернативной. Условие не выполняется

Отсутствие связей между остатками и предикторами

Для проверки этого условия нам понадобится построить диаграммы рассеивания для остатков и каждого предиктора. Начнём с длины слова - переменной **LengthInLetters**, визуализируем:

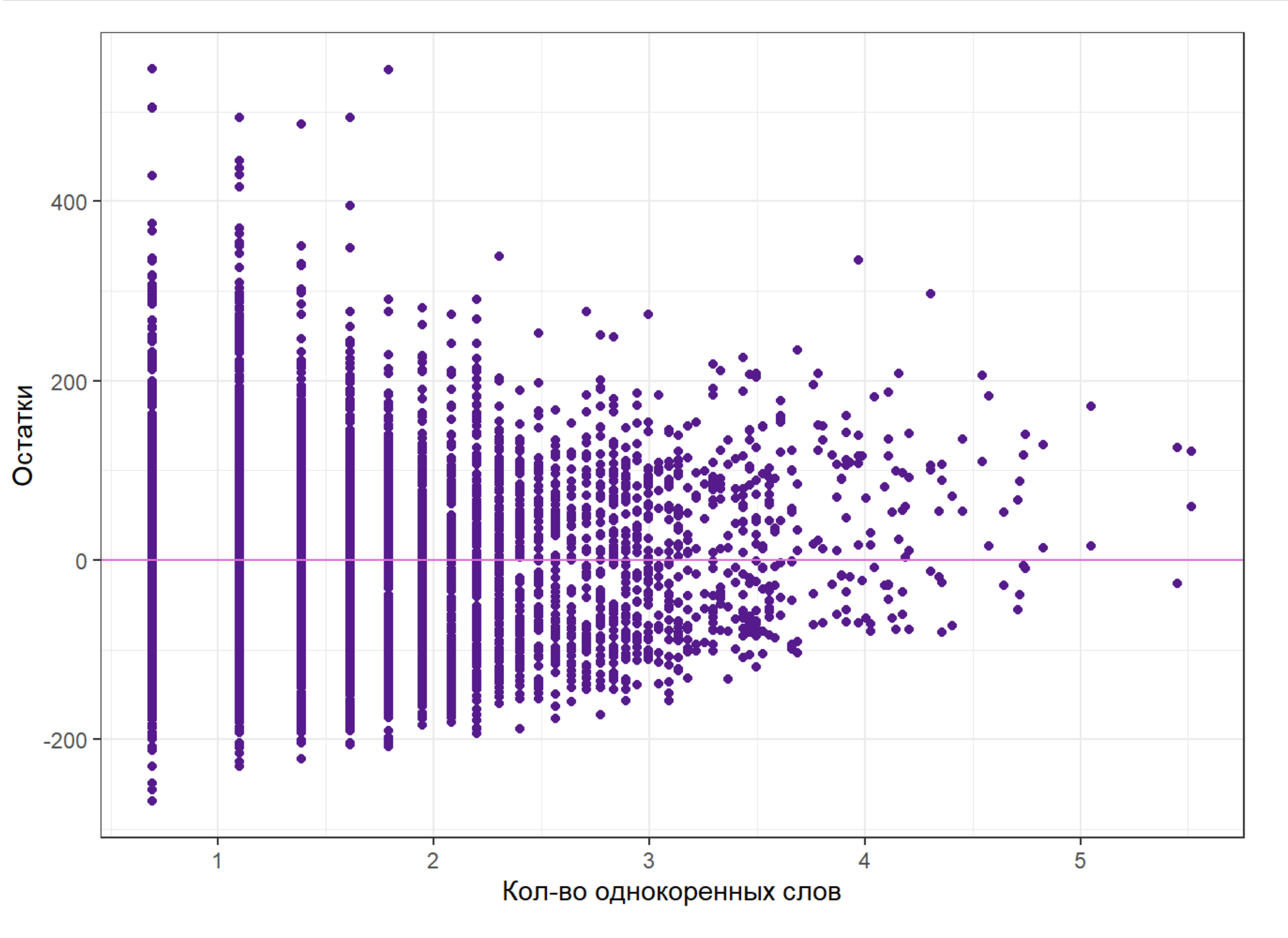
```
ggplot(data = df, aes(x = LengthInLetters, y = res))+
  geom_point() + geom_hline(yintercept = 0, color = "darkorange")+
  ylab("Остатки")+
  xlab("Кол-во букв в слове")+
  theme_bw()
```



Как мы видим, в левой и правой частях графика остатки заметно меньше. Наблюдается связь между количеством слов и остатков.

Проверим с переменной **FamilySize**, построим диаграмму:

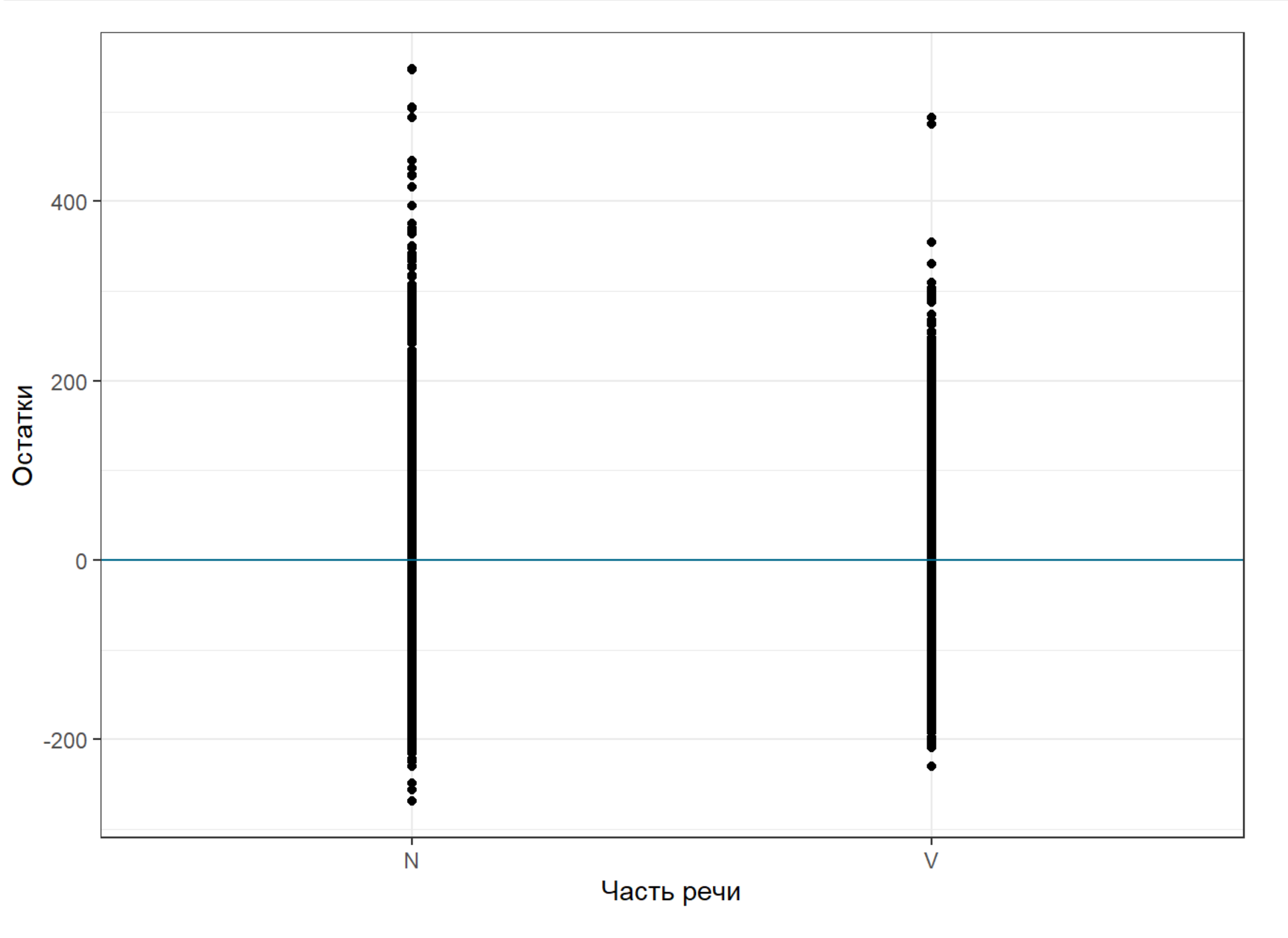
```
ggplot(data = df, aes(x = FamilySize, y = res))+
  geom_point(color = "purple4")+
  geom_hline(yintercept = 0, color = "orchid")+
  ylab("Остатки")+
  xlab("Кол-во однокоренных слов")+
  theme_bw()
```



Весьма заметно, что ошибки меньше с ростом занчения переменной, то есть с ростом количества однокоренных слов.

Проверим связь с переменной **WordCategory**, то есть с частью речи, построим диаграмму:

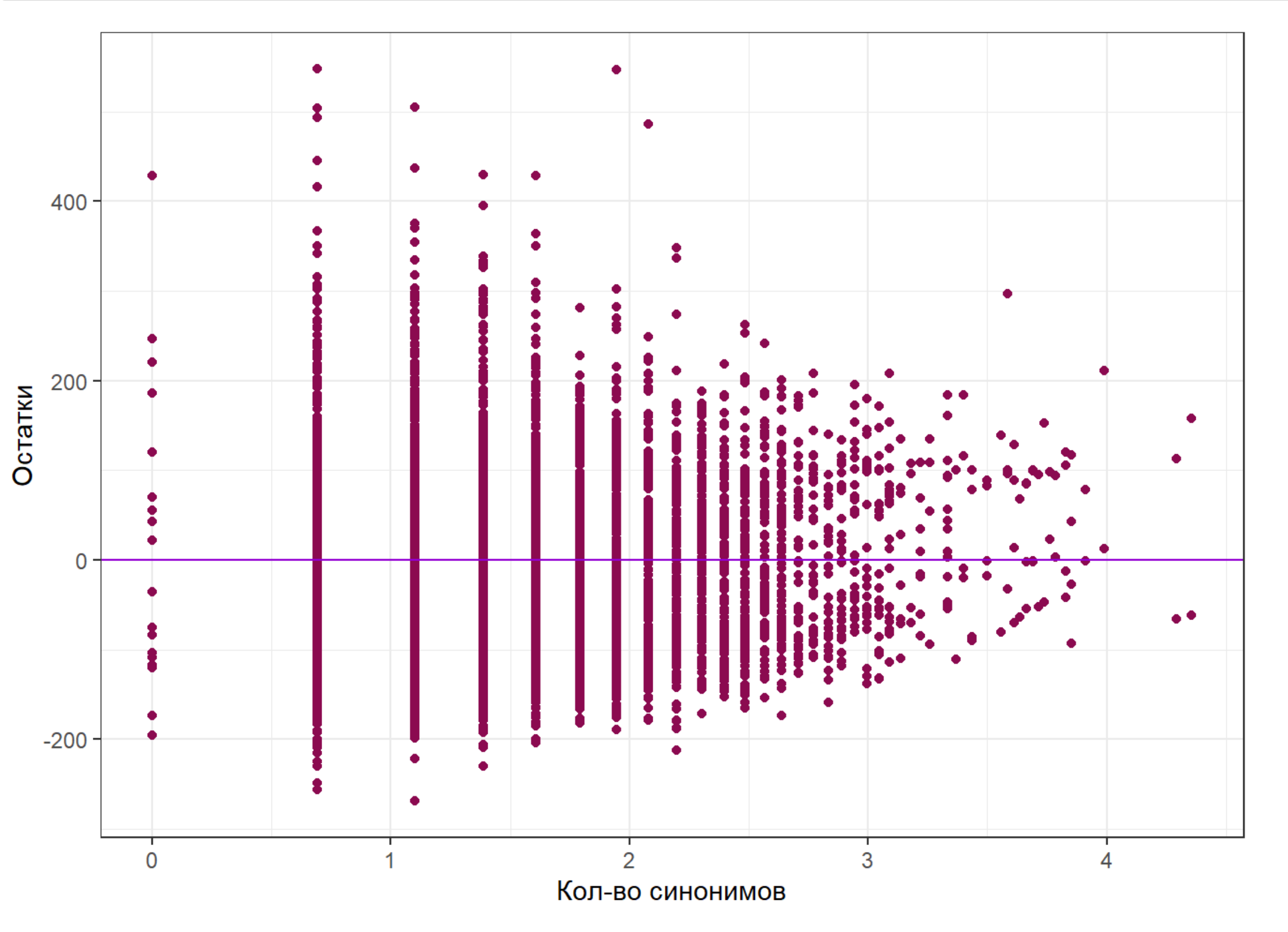
```
ggplot(data = df, aes(x = WordCategory, y = res))+
  geom_point()+
  geom_hline(yintercept = 0, color = "deepskyblue4")+
  ylab("Остатки")+
  xlab("Часть речи")+
  theme_bw()
```



Значения остатков так же заметно отличаются для существительных(N) и глаголов(V).

Проверим выполнения этого условия для переменной **NumberSimplexSynsets**, построим диаграмму для этого:

```
ggplot(data = df, aes(x = NumberSimplexSynsets, y = res))+
  geom_point(color = "deeppink4")+
  geom_hline(yintercept = 0, color = "darkviolet")+
  ylab("Остатки")+
  xlab("Кол-во синонимов")+
  theme_bw()
```



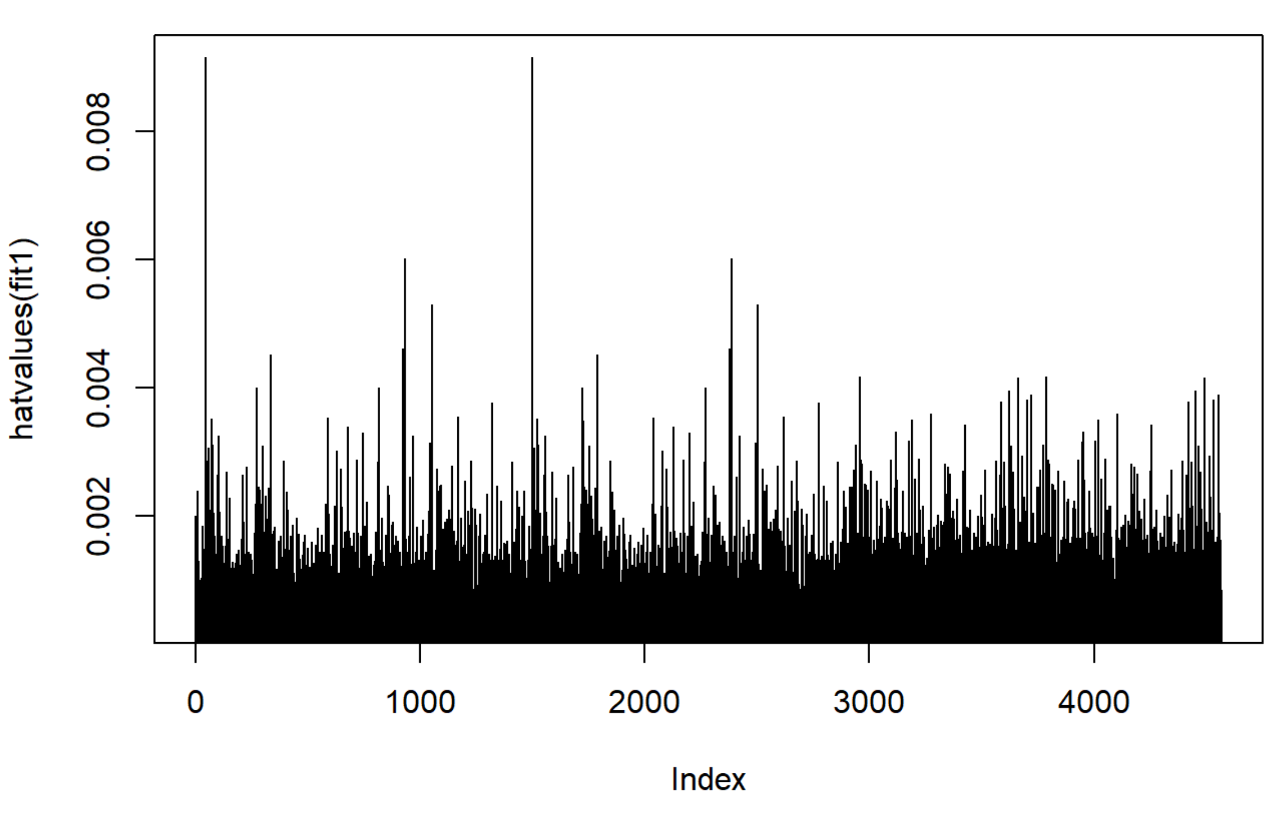
Остатки становятся заметно меньше с увеличением числа синонимов, связь с предиктором есть.

Между остатками и каждым предиктором есть связь. Условие не выполняется.

Наличие влиятельных наблюдений

Для проверки наличия влиятельных наблюдений построим соответствующую гистограмму:

```
plot(hatvalues(fit1), type = 'h')
```



Можем сказать, что влиятельные наблюдения отсутствуют.

Итоги

Из проделанного нами мини-исследования можем сделать следующие выводы. Богатство морфологической семьи искомого слова и количество синонимов у этого слова влияют на время, затраченное на его узнавание. Мы подтверждаем вторую и четвёртую гипотезы нашего мини-исследования.
Но к сожалению, предсказательная сила модели мала, $R^2 = 13,79\%$. Построенная модель объясняет лишь 13,39% изменчивости зависимой переменной. Также среди наших предикторов значимые коэффициенты имеют лишь два: **FamilySize** - количество однокоренных слов и **NumberSimplexSynsets** - количество слов-синонимов. Ни одно условие Гаусса-Маркова не выполняется, отсутствуют влиятельные наблюдения. В целом, наша модель, к сожалению, получилась неудачной, над этим стоит поработать в дальнейшем.