Федоров Макар

Описание данных

В файле `english.csv` сохранены результаты психолингвистического эксперимента, посвященного узнаванию и называнию слов (*lexical decision* и *word naming*). В ходе первой части эксперимента участники должны были решить, является ли слово, демонстрируемое на экране, реально существующим или нет. Другими словами, участники определяли, является ли слово настоящим или оно искусственно сконструировано по правилам грамматики. Во второй части эксперимента участники должны были прочитать показанное на экране слово вслух. В течение обеих частей эксперимента фиксировалось время реакции на слово: как быстро участник кликает на кнопки «слово» или «не-слово» (узнавание слов) или читает само слово (называние слов).

Основные показатели в файле:

- `AgeSubject`: возраст участника (`young` или `old`);
- `WordCategory`: часть речи слова (`N` — существительное, `V` — глагол);
- `RTlexdec`: время, затраченное на узнавание слова, в миллисекундах;
- `RTnaming`: время, затраченное на называние слова, в миллисекундах;
- `WrittenFrequency`: частота встречаемости слова в письменных текстах;
- `LengthInLetters`: длина слова, в буквах;
- `FamilySize`: логарифмированный размер морфологической семьи слова (количества однокоренных слов);
- `NumberSimplexSynsets`: логарифмированное количество синсетов (наборов синонимов) в WordNet, в которые входит слово.

Задача 1

Загрузите данные из файла `english.csv` и сохраните их в датафрейм `english`. Используя имеющиеся данные, постройте модель, которая объясняла бы, каким образом время, которое люди тратят на угадывание реалистичности слова (`RTlexdec`), зависит от длины этого слова, встречаемости этого слова в письменных текстах и количества синонимов. Проинтерпретируйте полученные результаты. Ваша интерпретация должна включать следующие элементы:

- указание на то, коэффициенты при каких переменных являются статистически значимыми (и на каком уровне значимости);
- объяснения, каким образом, в среднем, изменяется зависимая переменная при изменении независимых на единицу.

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.1.2     ✓ dplyr   1.0.6
```

```
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.1

## ── Conflicts ─────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
english <- read_csv("english.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

##
## ── Column specification
──────────────────────────────────────────────────────────────────────
## cols(
##   .default = col_double(),
##   Word = col_character(),
##   AgeSubject = col_character(),
##   WordCategory = col_character(),
##   CV = col_character(),
##   Obstruent = col_character(),
##   Frication = col_character(),
##   Voice = col_character()
## )
## i Use `spec()` for the full column specifications.
```

```
model <- lm(data = english, RTlexdec ~ LengthInLetters + WrittenFrequency +
NumberSimplexSynsets)

summary(model)
```

```
##
## Call:
## lm(formula = RTlexdec ~ LengthInLetters + WrittenFrequency +
##     NumberSimplexSynsets, data = english)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -306.26  -82.90   -7.46   70.15  549.92
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           843.062      9.423  89.468  < 2e-16 ***
## LengthInLetters         2.763      1.818   1.520    0.129
## WrittenFrequency      -23.385      1.001 -23.367  < 2e-16 ***
## NumberSimplexSynsets  -17.278      2.750  -6.282 3.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.1 on 4564 degrees of freedom
```

```
## Multiple R-squared:  0.1947, Adjusted R-squared:  0.1942
## F-statistic: 367.8 on 3 and 4564 DF,  p-value: < 2.2e-16
```

Уравнение прямой получается следуюшим: RTlexdec = 843.062 -23.385 * WrittenFrequency -17.278 * NumberSimplexSynsets + 2.763 * LengthInLetters
Коэффициенты при встречаемости слова в текстах и количестве синонимов являются статистически значимыми при уровне значимости 0.001.Длина слова в буквах статистически значимой не являеется (если мы не берем уровень значимости > 0.129). То есть,при увеличении частоты встремаемости слова на одну единицу, скорость узнавание в среднем становится меньше на 23.385 миллисекунды. При увеличении логарифмированного количества синонимов на одну единицу, скорость узнавание в среднем становится меньше на 17.278 миллисекунд. При увеличении длины слова на 1, скорость узнавания увеличивается на 2.763 (при этом этот показатель не является статистически значимым).

Задача 2

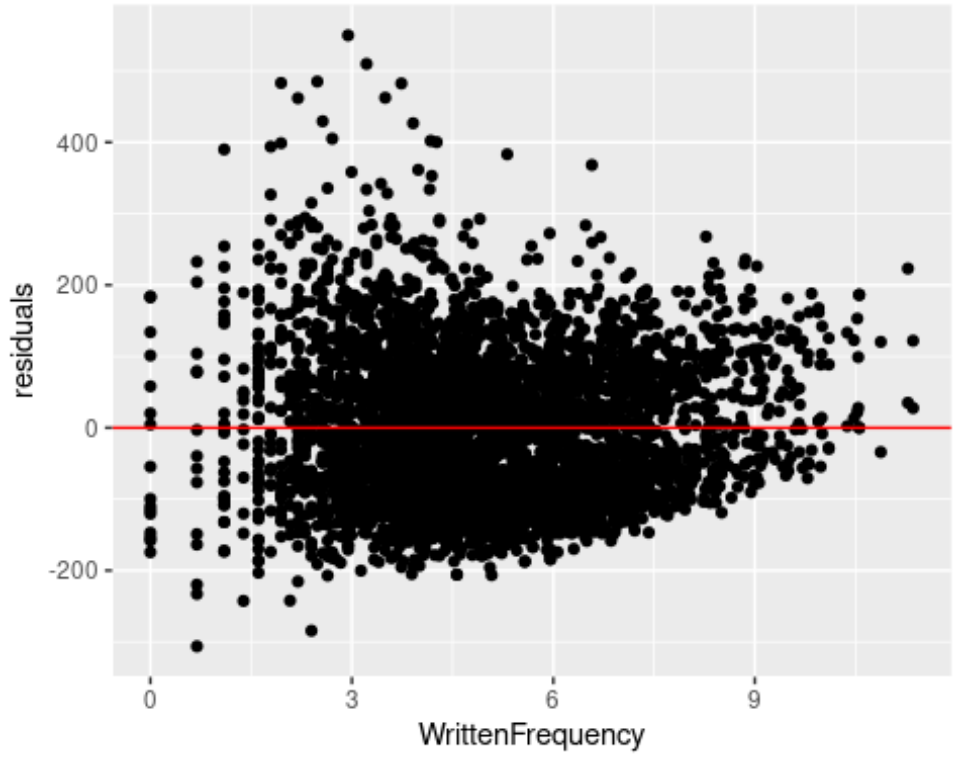
Проведите исследования качества полученной модели: проверьте её на наличие мультиколлинеарности, гетероскедастичности и влиятельных наблюдений. Если будут выявлены проблемы, устраните их (пересчитайте коэффициенты, устойчивые к гетероскедастичности, удалите влиятельные наблюдения и оцените модель ещё раз). Если после корректировки модели произошли существенные изменения, поясните, в чём они заключаются.

```
# Для того, чтобы проверить на мультиколлинеарность, можно построить матрицу корреляции
cor(english[, c(2, 8, 13, 15)])
```

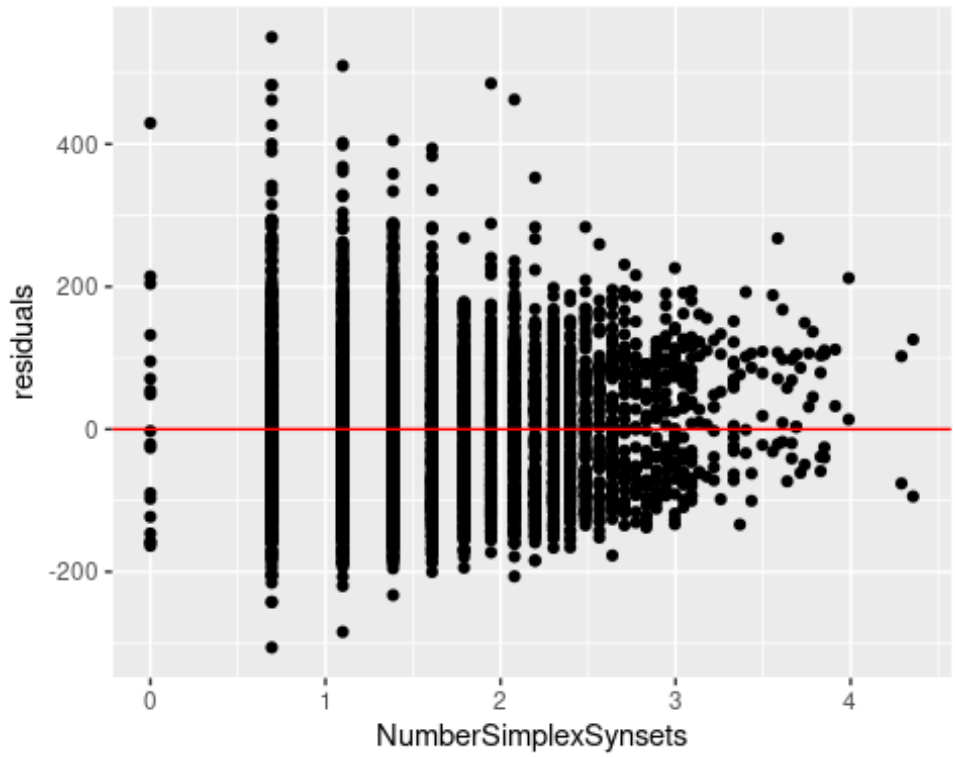
```
##                          RTlexdec WrittenFrequency NumberSimplexSynsets
## RTlexdec               1.00000000      -0.43295220         -0.310537129
## WrittenFrequency      -0.43295220       1.00000000          0.558749584
## NumberSimplexSynsets -0.31053713       0.55874958          1.000000000
## LengthInLetters       0.04590441      -0.06663196         -0.006364411
##                       LengthInLetters
## RTlexdec                   0.045904412
## WrittenFrequency          -0.066631955
## NumberSimplexSynsets      -0.006364411
## LengthInLetters            1.000000000
```

```
# Наибольшая корреляция - между WrittenFrequency и NumberSimplexSynsets = 0.55874958. Это, однако, в пределах нормы, поэтому ни один из параметров мы удалять из модели не будем

# Для проверки на наличие гетероскедастичности можно построить диаграмму рассеивания «независимая переменная vs ошибки"
library(sandwich)
library(lmtest)
```

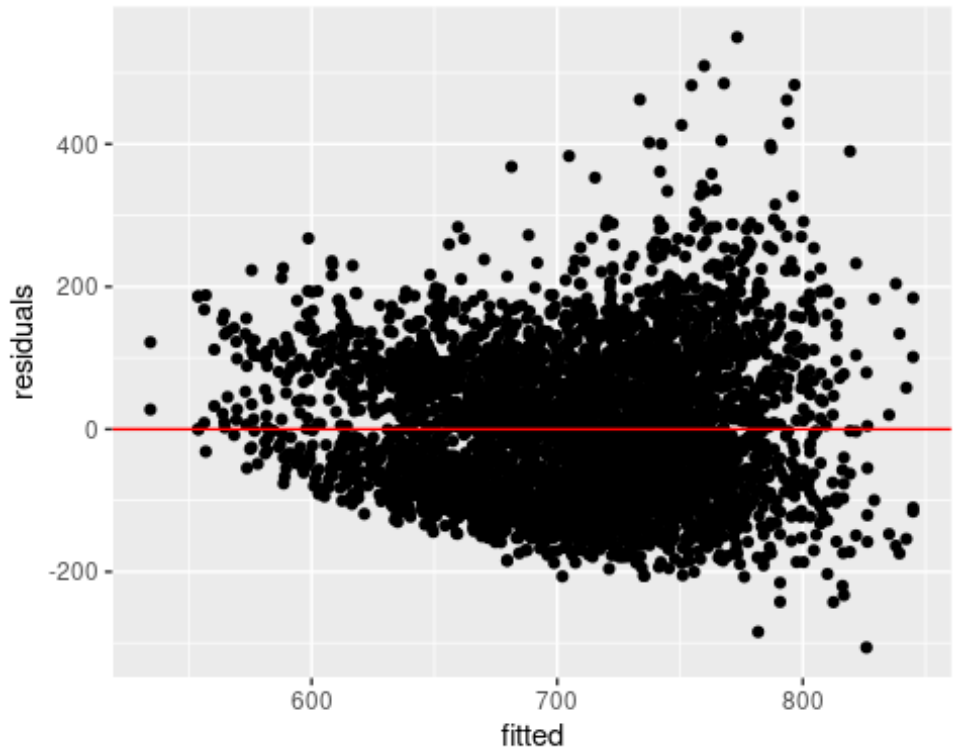
```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

english$residuals <- model$residuals
english$fitted <- model$fitted.values
ggplot(data = english, aes(x = WrittenFrequency, y = residuals)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```



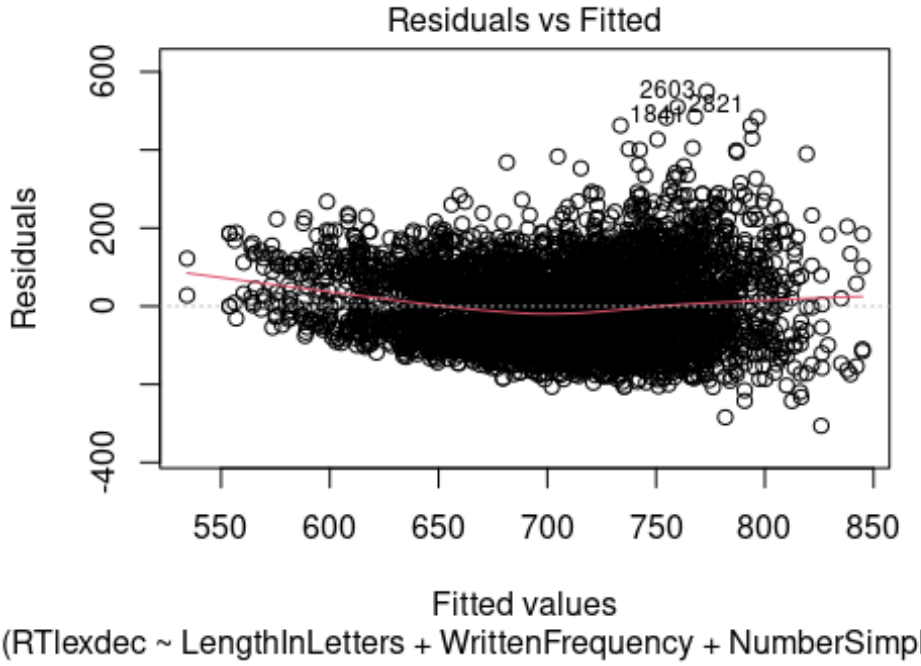

```
# Видно, что для низких зачений WrittenFrequency ошибок больще
ggplot(data = english, aes(x = NumberSimplexSynsets, y = residuals)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```

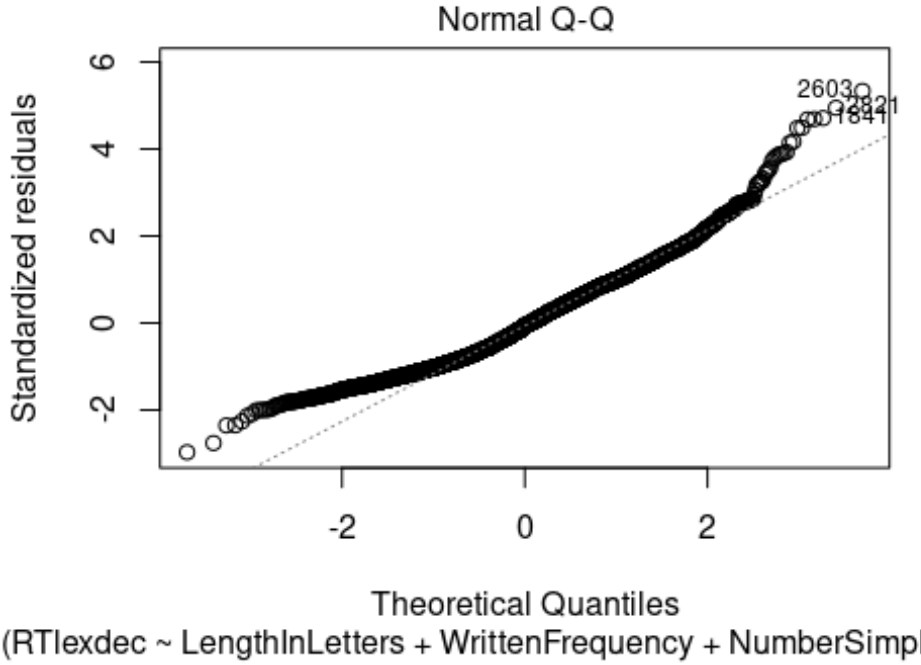


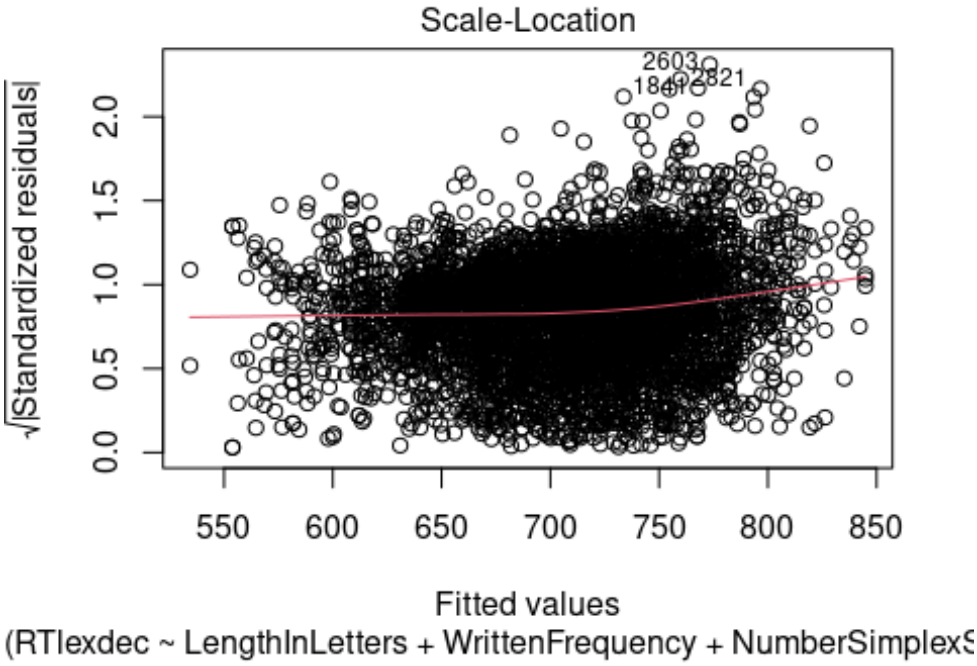

```
# Здесь видно, что разброс больше на небольших значениях
# Для всех переменных можно построить диаграмму рассеивания  «предсказанные значения vs ошибки"
ggplot(data = english, aes(x = fitted, y = residuals)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```

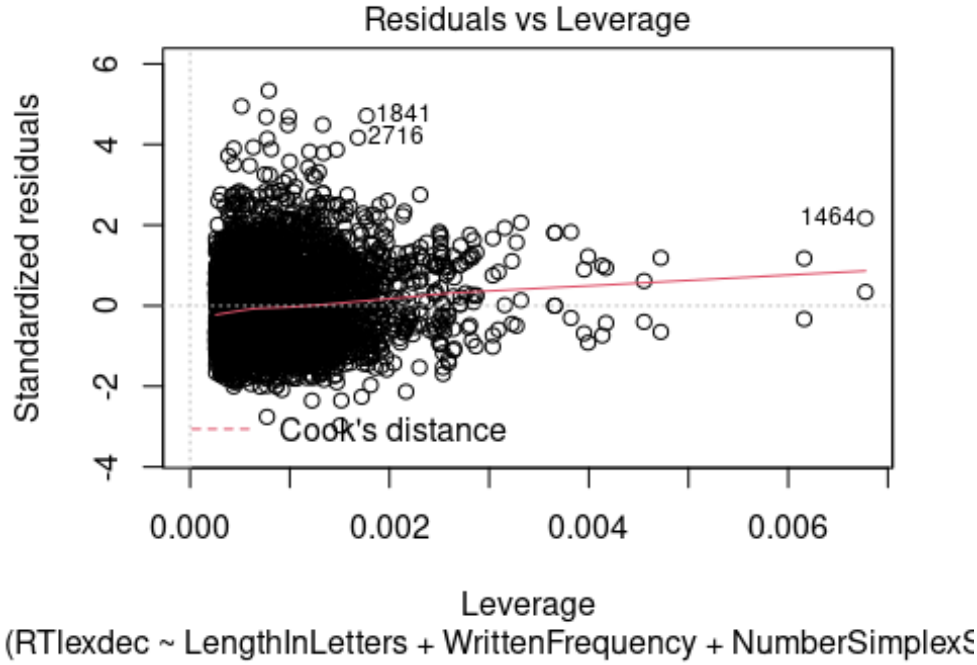



```
# Здесь виден паттерн, которого не должно быть - чем больше значения по оси x, тем больше разброс значений по оси y
# Для того, чтобы решить проблему гетероскедастичности,перерасчитаем с импользованием heteroscedasticity-consistent standard errors
coeftest(model, vcov = vcovHC(model, type = "HC0"))

##
## t test of coefficients:
##
##                       Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)          843.0618    10.0658  83.7551 < 2.2e-16 ***
## LengthInLetters        2.7634     1.8602   1.4855    0.1375
## WrittenFrequency     -23.3853     1.0401 -22.4840 < 2.2e-16 ***
## NumberSimplexSynsets -17.2779     2.7483  -6.2867 3.546e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Изменилась стандартная ошибка, t-value и  p-value, но не сильно.
# Теперь проверим модель на наличие влиятельных наблюдений. Для этого используем график "residuals vs leverage"
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
(RTlexdec ~ LengthInLetters + WrittenFrequency + NumberSimplexS
2603
2821
1841
600
200
0
-400
550
600
650
700
750
800
850


Normal Q-Q
Standardized residuals
Theoretical Quantiles
(RTlexdec ~ LengthInLetters + WrittenFrequency + NumberSimplexS
2603
2821
1841
6
4
2
0
-2
-2
0
2






```
# Наблюдений с высокой влиятельностью и высокими ошибками на графике нет
(они были бы помечены с использованием "Cook's distance")
```

Задача 3

Обновите модель из задачи 2 в предположении, что на время, которое люди тратят на угадывание реалистичности слова, также влияет часть речи, и при этом влияние длины слова на это время не одинаковое для существительных и глаголов. Приведите ответы (с обоснованием) на следующие вопросы.

1. Можно ли сказать, что, в среднем, люди более быстро угадывают существительные, чем глаголы?

2. Можно ли сказать, что, длина слова оказывает значимо разный эффект на время, которое люди тратят на угадывание, в случаях, если это слово является существительным и глаголом?

```
model2 <- lm(data = english, RTlexdec ~ LengthInLetters + WrittenFrequency +
NumberSimplexSynsets + WordCategory + LengthInLetters:WordCategory)
summary(model2)
```

```
##
## Call:
## lm(formula = RTlexdec ~ LengthInLetters + WrittenFrequency +
##     NumberSimplexSynsets + WordCategory + LengthInLetters:WordCategory,
##     data = english)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -306.97  -82.95   -7.64   70.54  549.73
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     833.429     11.094  75.122  < 2e-16 ***
## LengthInLetters                   5.247      2.317   2.265   0.0236 *
## WrittenFrequency                -23.617      1.012 -23.338  < 2e-16 ***
## NumberSimplexSynsets            -16.456      2.861  -5.752 9.38e-09 ***
## WordCategoryV                    23.231     16.784   1.384   0.1664
## LengthInLetters:WordCategoryV    -6.084      3.752  -1.621   0.1050
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.1 on 4562 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1945
## F-statistic: 221.5 on 5 and 4562 DF,  p-value: < 2.2e-16
```

Уравнение модели получается следующим: RTlexdec = 833.429 -23.617 * WrittenFrequency + 5.247 * LengthInLetters -16.456 * NumberSimplexSynsets - 6.084 * LengthInLetters * WOrdCategory + 23.231 + WOrdCategory. WOrdCategory = V принимает значение 1,WOrdCategory = N принимает значение 0 1) Влияние части речи на скорость угадывания не является статистически значимой, p-value = 0.1664. В среднем, глаголы увеличивает время на 23.231 миллисекунды, но этот коэффициент при этом параметре не

является статистически значимым, поэтому нельзя сделать обоснованный вывод, что люди быстрее угадывают существительные. 2) Влияние длины слова для разных частей речи также не является статистически значимой, p-value = 0.1050. Если слово является глаголом, то при увеличении длины слова на одну единицу, время, затрачиваемое на угадывание слова уменьшается на 6.084. Однако, обоснованно заключить, что длина слова оказывает более значимый эффект в зависимости от части речи нельзя, так коэффициент при этом параметре не является статистически значимым.