The use of expressive suffixes in common and proper nouns in Russian

Vakhranyov

26 04 2020

##A hypothesis formulated in terms of subject

In this research we are going to check, if expressive suffixes of the same number of letters are more often used in the proper nouns in Russian than in the common nouns. In the work of K.M. Medvedeva we can find the statement that “the variety of word-building models is one of the features of proper names: “A feature of personal names is that they have a great ability to form variants, or derivatives. “Derivatives unite all derivative names: shorten, affectionate, diminutive and familiar <…>, those that can’t be clearly differentiated”". So we can inerpret this as the statement, in which it is said that proper nouns are very flexible in creating new forms, including those that were made by the use of evpressive suffixes. Also we a going to check, if their regularity of use in those two groups of nouns is dependent on some extra data (the order of the suffix and the grammatical gender of the noun).

##Research design

So we need a dataset of Russian nouns, where would be collected both — proper and common. We should mark up several features in this dataset, to check if they are significant for our hypothesis: the number of letters in the suffix, the order of the suffix in the word (if it is the first suffix, the second or the third) and the grammatical gender of the noun, where that particular suffix is used.

Our null hypothesis would be formulated in a such way: there is no difference in the regularity of use of expressive suffixes of any number of letters (for those, that are used with nouns) in common and proper nouns. The alternative: there is a difference in the regularity of use of expressive suffixes of any number of letters in common and proper nouns.

For our research we are going to use the model of the linear regression.

##Description of the data collection method

The data would is collected for the research in three ways: 1) by a search of the nouns with the expressive suffixes in The General Internet Corpus of Russian; 2) by a search in the internet articles, dedicated to the expressve suffixes in Russia: “Эмоционально-экспрессивная окраска слов” — Розенталь Д.Э., Голуб И.Б., Теленкова М.А., “Современный русский язык” “Значение суффиксов для русской литературы” — http://antisochinenie.ru/сочинения/Другие_сочинения_по_зарубежной_литературе/Значение_суффиксов_для_русской_литературы (http://antisochinenie.ru/сочинения/Другие_сочинения_по_зарубежной_литературе/Значение_суффиксов_для_русской_литературы) “Эмоционально-экспрессивная окраска слов” — Саримова Р.Р., “Русский язык и культура речи” “Семантика эмоционально-экспрессивных суффиксов квалитативных форм русских антропонимов” — Медведева К. М., Молодой ученый, 2013, № 7 (54) “Русские суффиксы” — http://www.slovorod.ru/russian-suffixes.html (http://www.slovorod.ru/russian-suffixes.html) “Семантико-грамматическая классификация оценочных слов в рассказах М. Зощенко” — С. А. Пучкова “Оценочность суффиксов в русской речи” — В. Покровский “Экспрессивные диминутивы в условиях конкуренции с нейтральными существительными (на материале русского языка)” — И. Фуфаева; 3) by a search in the Russian grammatical dictionary by Zaliznyak.

##Collected data, their description

The dataset includes one csv-file. The table consists of information about 485 usage examples of the expressive suffixes in Russian nouns with the definition of the noun group (proper or common), number of letters in suffixes, grammatical gender and order of the suffixes in a word. Here are the columns:

- word — an example of a Russian word with one or more expressive suffix
- morpheme — an expressive suffix, that is observed in a particular row
- proper/common — here is the information about the fact, if a noun is proper or common
- order — the information about the order of a suffix in a word: is it the first suffix, the second or the third
- number of letters — how many letters in a particular suffix
- gender — the information about the grammatical gender of a noun

Several nouns in the column “word” are used twice. The reason for this is that we have in these words two suffixes, that can be marked as expressive.

Speaking about the order — it doesn’t matter if the other suffixes are not expressive, when we mark any expressive suffix as “second” or “third”: we count all suffies.

There also five marks for the different grammatical gender: “c” for “common gender”, “f” for “feminine”, “m” for “masculine”, “n” for “neuter gender” and “pl” for “pluralia tantum”.

1
2
3
4

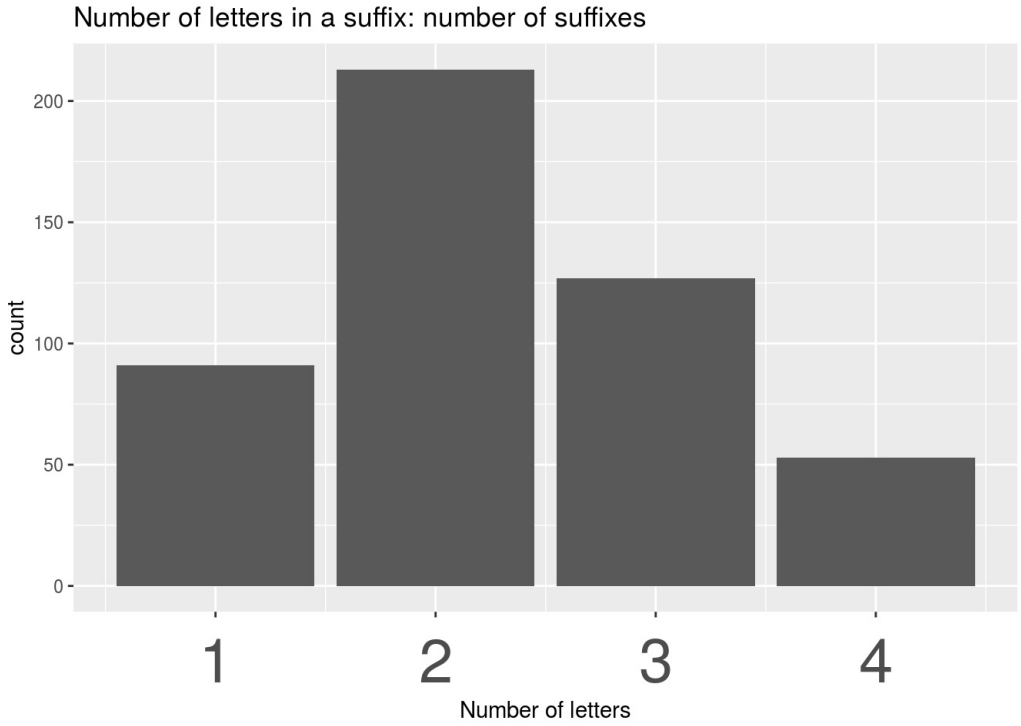
5
6

6 rows | 1-1 of 7 columns

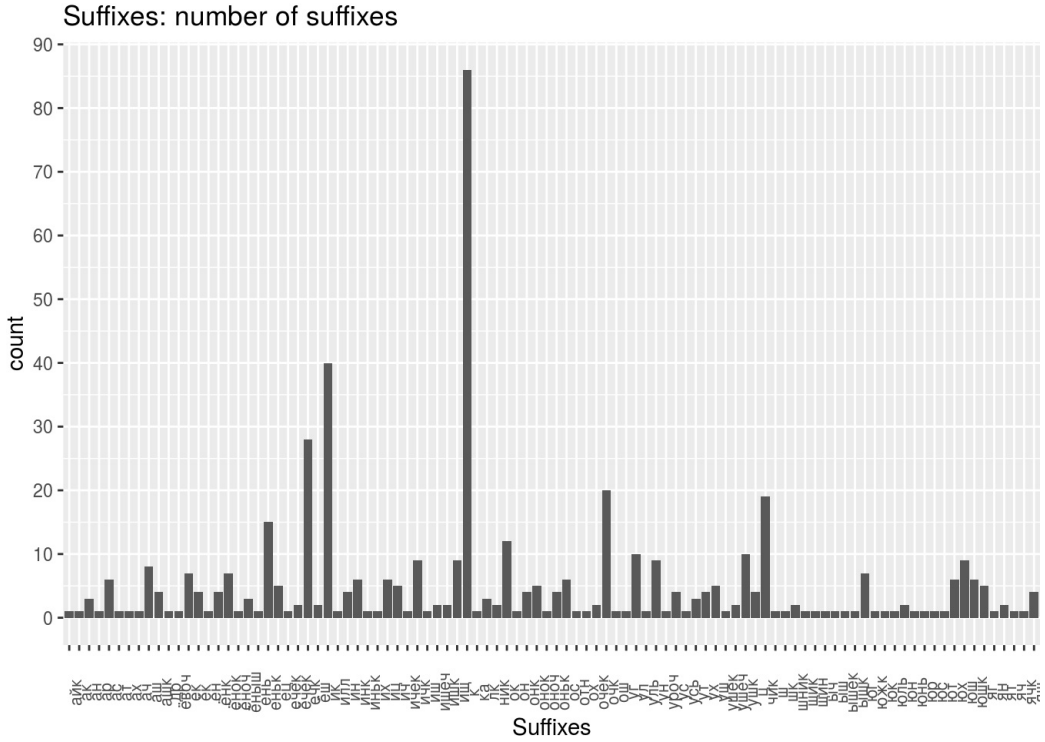
##Exploratory data analysis, descriptive statistic and vizualization

On the bar charts below quantities of a number of letters in suffixes, suffixes, types of nouns, grammatical genders and orders of suffixes are represented.

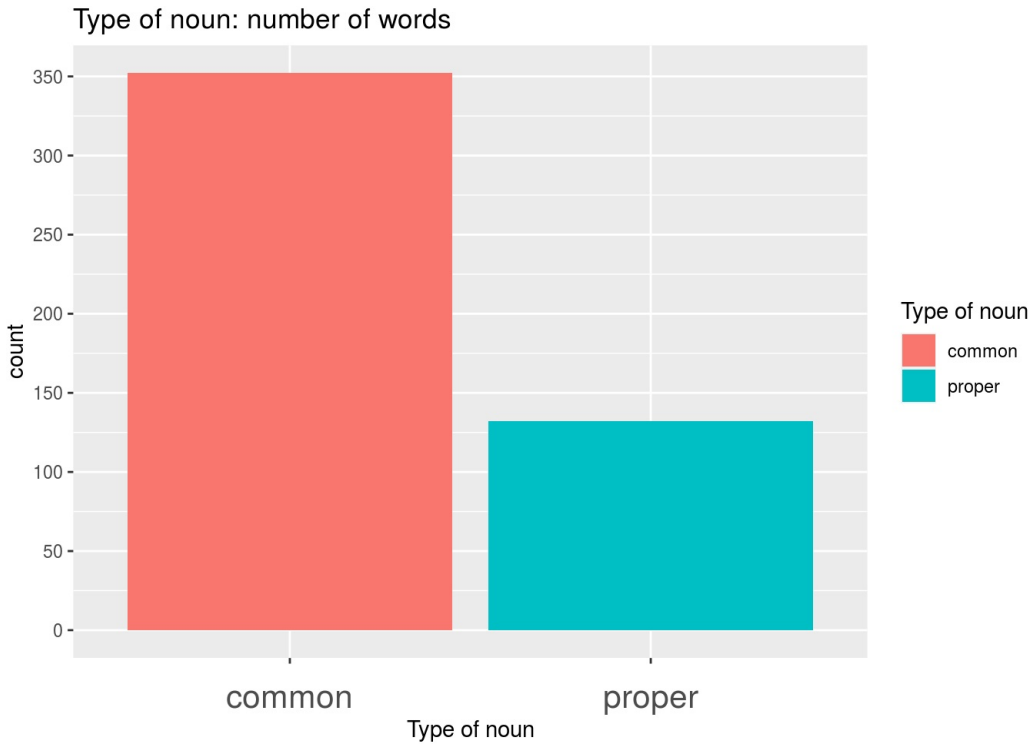
```
data %>%
ggplot(aes(number.of.letters, fill=number.of.letters)) + geom_bar() + labs(title="Number of letters in a suffix: n
umber of suffixes", x="Number of letters") + theme(axis.text.x=element_text(size=30, angle=0, margin=margin(t=10))
) + scale_y_continuous(breaks=seq(0,250,50))
```



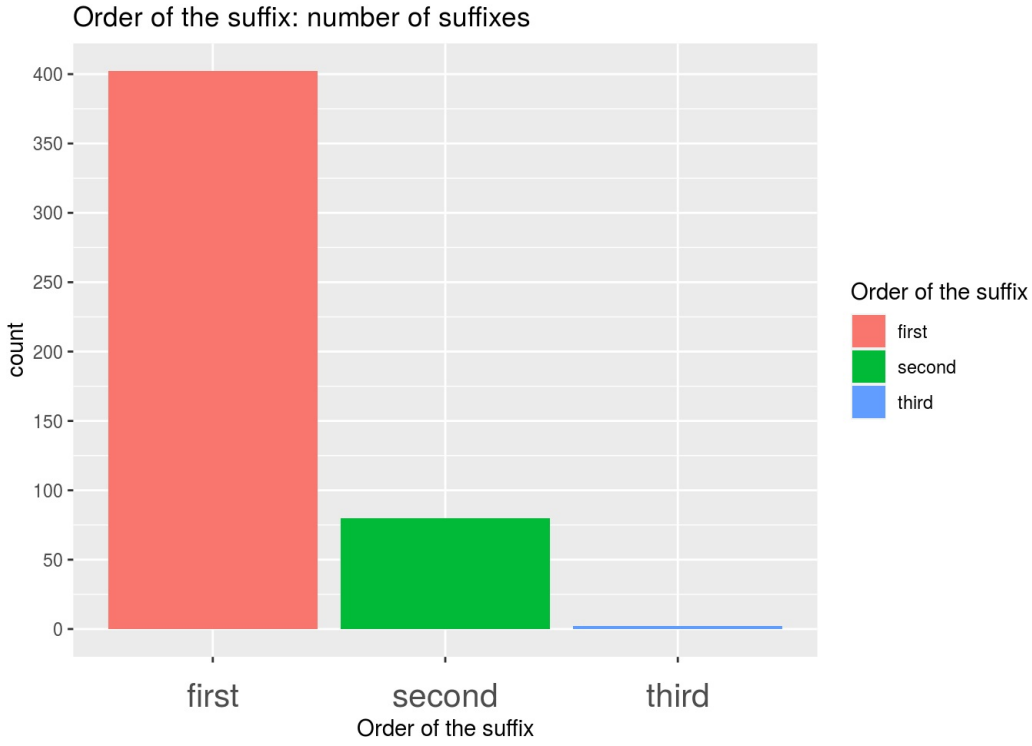

```
par(mar=c(7,7,1,1)+3.5,mgp=c(6,1,0))
data %>%
ggplot(aes(morpheme)) + geom_bar() + labs(title="Suffixes: number of suffixes", x="Suffixes") + theme(plot.margin=
margin(1, 0, 1, 0)) + theme(axis.text.x=element_text(size=8, angle=90, margin=margin(t=10))) + scale_y_continuous(
breaks=seq(0,250,10))
```



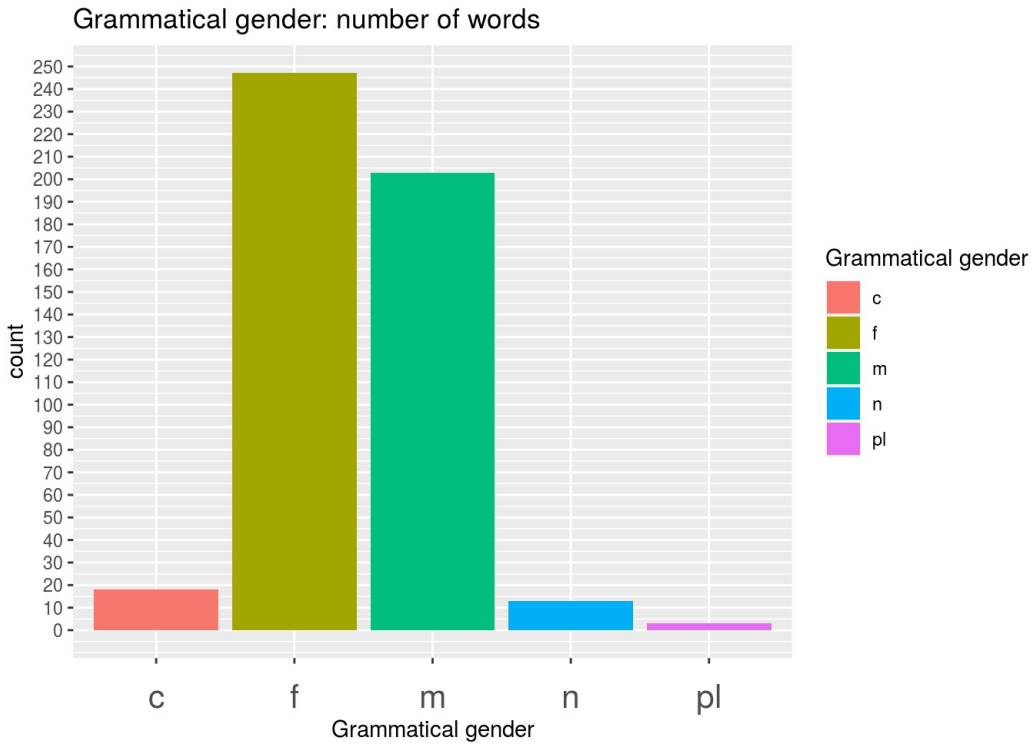
```
par(mar=c(7,7,1,1)+3.5,mgp=c(6,1,0))
data %>%
ggplot(aes(proper.common, fill=proper.common)) + geom_bar() + labs(title="Type of noun: number of words", x="Type
of noun") + scale_fill_discrete("Type of noun") + theme(plot.margin=margin(1, 0, 1, 0)) + theme(axis.text.x=elemen
t_text(size=16, angle=0, margin=margin(t=10))) + scale_y_continuous(breaks=seq(0,400,50))
```




```
par(mar=c(7,7,1,1)+3.5,mgp=c(6,1,0))
data %>%
ggplot(aes(order, fill=order)) + geom_bar() + labs(title="Order of the suffix: number of suffixes", x="Order of th
e suffix") + scale_fill_discrete("Order of the suffix") + theme(plot.margin=margin(1, 0, 1, 0)) + theme(axis.text.
x=element_text(size=16, angle=0, margin=margin(t=10))) + scale_y_continuous(breaks=seq(0,400,50))
```

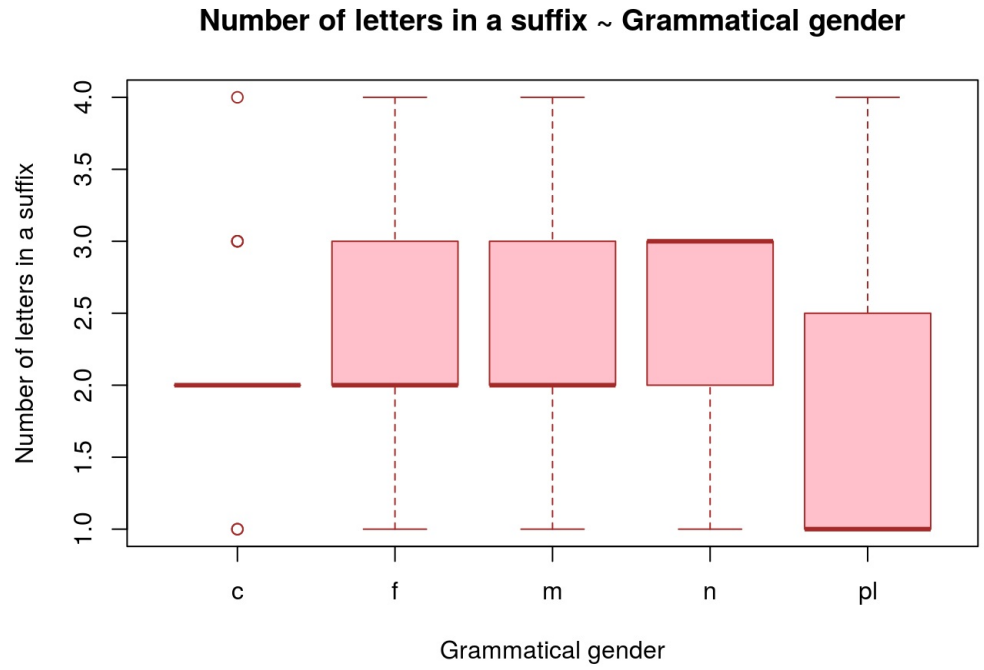



```
par(mar=c(7,7,1,1)+3.5,mgp=c(6,1,0))
data %>%
ggplot(aes(gender, fill=gender)) + geom_bar() + labs(title="Grammatical gender: number of words", x="Grammatical g
ender") + scale_fill_discrete("Grammatical gender") + theme(plot.margin=margin(1, 0, 1, 0)) + theme(axis.text.x=el
ement_text(size=16, angle=0, margin=margin(t=10))) + scale_y_continuous(breaks=seq(0,250,10))
```



We are able to notice some outliers in the dataset: 12 out of 18 nouns of the common gender have two-letter expressive suffixes.

```
length(boxplot(data$number.of.letters ~ data$gender,
        main = "Number of letters in a suffix ~ Grammatical gender",
        xlab = "Grammatical gender",
        ylab = "Number of letters in a suffix",
        col = "pink",
        border = "brown")$out)
```

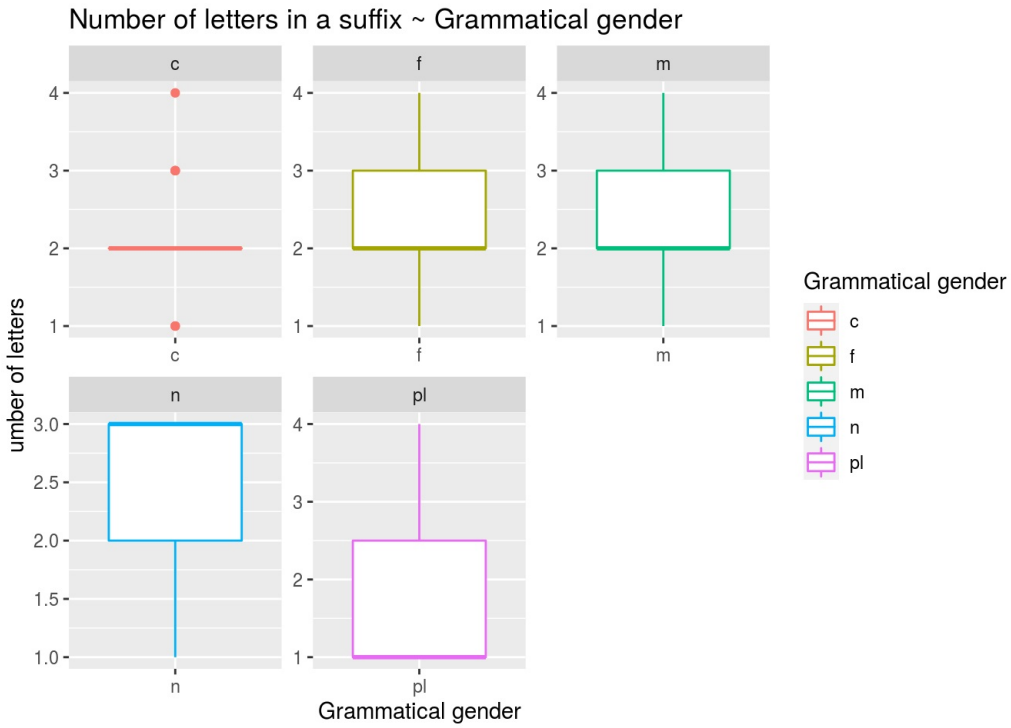



```
## [1] 6
```

```
# [1] 32
```

The boxplots below demonstrate the same results, but separated by genders.

```
p <- ggplot(data, aes(gender, number.of.letters, color=gender))
p + geom_boxplot() + facet_wrap(~gender, scale="free") + labs(title = "Number of letters in a suffix ~ Grammatical
gender", x = "Grammatical gender", y = "umber of letters") + scale_colour_discrete("Grammatical gender")
```



##Results of applying statistical tests

Here we create a column of zeros and ones for proper and common nouns, to make this information usefull for our logistic regression model.

```
data$bin_nouns[data$proper.common=="proper"] <- 1
data$bin_nouns[data$proper.common=="common"] <- 0

head(data)
```

1
2
3
4
5
6

6 rows | 1-1 of 8 columns

The code below estimates a logistic regression model using the glm (generalized linear model) function. First, we convert number.of.letters, order and gender to factors to indicate that they should be treated as categorical variables.

```
data$number.of.letters <- factor(data$number.of.letters)
data$order <- factor(data$order)
data$gender <- factor(data$gender)
mylogit <- glm(bin_nouns ~ number.of.letters + order + gender, data = data, family = 'binomial')
```

Since we gave our model a name (mylogit), R will not produce any output from our regression. In order to get the results we use the summary command:

```
summary(mylogit)
```

```
##
## Call:
## glm(formula = bin_nouns ~ number.of.letters + order + gender,
##     family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0147  -0.8131  -0.7419   1.4692   1.9094
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.4226     0.6205  -2.293   0.0219 *
## number.of.letters2    0.2732     0.2910   0.939   0.3478
## number.of.letters3   -0.2240     0.3243  -0.691   0.4897
## number.of.letters4    0.2678     0.3902   0.686   0.4925
## ordersecond           0.2686     0.2727   0.985   0.3247
## orderthird          -14.5949  1028.9564  -0.014   0.9887
## genderf               0.4853     0.5913   0.821   0.4117
## genderm               0.1437     0.5940   0.242   0.8089
## gendern              -0.2662     0.9642  -0.276   0.7825
## genderpl            -14.4118   840.2744  -0.017   0.9863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 567.20  on 483  degrees of freedom
## Residual deviance: 556.42  on 474  degrees of freedom
## AIC: 576.42
##
## Number of Fisher Scoring iterations: 14
```

In the output above, the first thing we see is the call, this is R reminding us what the model we ran was, what options we specified, etc.

Next we see the deviance residuals, which are a measure of model fit. This part of output shows the distribution of the deviance residuals for individual cases used in the model.

The next part of the output shows the coefficients, their standard errors, the z-statistic (also known as a Wald z-statistic), and the associated p-values. All of them are statistically insignificant. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

For example, having the two-letter suffix, versus a one-letter suffix, changes the log odds of a noun being proper by 0.273.

Below the table of coefficients are fit indices, including the null and deviance residuals and the AIC.

##The results of the study

We have found, that there is no specific difference in the suffixes of common and proper nouns. The number of letters, grammatical gender and the order are not really significant for the type of noun.