Final Project of LING-411

Mehmet Akif Güçyener

Last edited 2025-01-05

Contents

Objective and Overview of Dataset 1

Lexical Decision and Naming Dataset 2

Data Analysis and Hypothesis Forming 2

Checking for Duplicates and NA's 2

Data Exploration 3

Hypothesis 4

Null Hypothesis 4

Alternative Hypothesis 4

Testing the Hypothesis 4

Transformations 4

Models 5

Model-1 5

Model-2 6

Model-3 6

Model-3_2 6

Model-7 7

Model Interpretation and Conclutions: 7

Objective and Overview of Dataset

This is a data analysis project. I have chosen one of the three datasets that was provided. Step-by-step we will try to understand the data, form a hypothesis around it and then test it using regression models.

Sidenote: I always wondered how you built the class notes for the course. In the last PS of the class Onur showed us how and I tried to give it a shot. Many thanks to both of you to boost our creativity.

Lexical Decision and Naming Dataset

The `eng_lexdec_naming` dataset contains data of a lexical decision and lexical naming task. In a decision task, participants try to determine whether a string is a word or not. Naming tasks involves showing participants pictures and asking them to name the picture. Reaction times are in miliseconds.

```
## # A tibble: 4,568 x 10
##    Word  RTlexdec RTnaming Familiarity WrittenFrequency FamilySize NounFrequency
##    <chr>    <dbl>    <dbl>       <dbl>            <dbl>      <dbl>         <dbl>
##  1 doe       695.     466.        2.37             50.0       4               49
##  2 stre~     547.     466.        5.6             669.        5              565
##  3 pork      617.     460.        3.87            151.        7.00           150
##  4 plug      633.     492.        3.93            133         9              170
##  5 prop      687.     477.        3.27            118.        4              125
##  6 dawn      584.     457.        3.73            592.        3              582
##  7 dog       527.     444.        5.67           1286.       40.0           2061
##  8 arc       741.     454.        3.1             133         5              144
##  9 skirt     536.     483.        4.43            376.        6.00           522
## 10 spree     740.     491.        3.27             30.0       2               32
## # i 4,558 more rows
## # i 3 more variables: VerbFrequency <dbl>, Voice <chr>, AgeSubject <chr>
```

It seems that our data consist of 4568 datapoints and 10 variables with column names Word, RTlexdec, RTnaming, Familiarity, WrittenFrequency, FamilySize, NounFrequency, VerbFrequency, Voice, AgeSubject. Inspection of the dataset shows that our Word column has a factor with two levels saved as AgeSubject: young, old. This makes the same word appear twice.

Data Analysis and Hypothesis Forming

Checking for Duplicates and NA's

In order to form a hypothesis, we must have an understanding of what the data is saying. Before that, it is good practice to check for duplicates and missing values. Not accounting for them may disrupt our analysis.

```r
sum(is.na(eng_lexdec_naming))
```

```
## [1] 0
```

```r
length(unique(eng_lexdec_naming$Word))
```

```
## [1] 2197
```

It seems that there is no missing values but there are some duplicates. Our word count had to be 2284, half of 4568 (total row number). We are missing 87 rows. I tried to check for duplicates by filtering words that may be appearing more than two times. I found out that some words like "arm" appear more than what they are supposed to:

Table 1: Table: Duplicate Word ‘Arm’

Word	RTlexdec	RTnaming	Familiarity	WrittenFrequency	FamilySize	NounFrequency	VerbFrequency	Voice	AgeSubject
arm	532.21	456.9	5.23	1897	39	3766	100	voiced	young
arm	681.21	645.2	5.23	1897	39	3766	100	voiced	old
arm	532.21	456.9	5.23	1897	39	3766	100	voiced	young
arm	681.21	645.2	5.23	1897	39	3766	100	voiced	old

As we can see, there are two duplicates. Each one comes from one AgeSubject. One is young, the other is old. From this point on, I have decided to get rid of them.

```r
eng_lexdec_naming %<>%
  distinct()
```

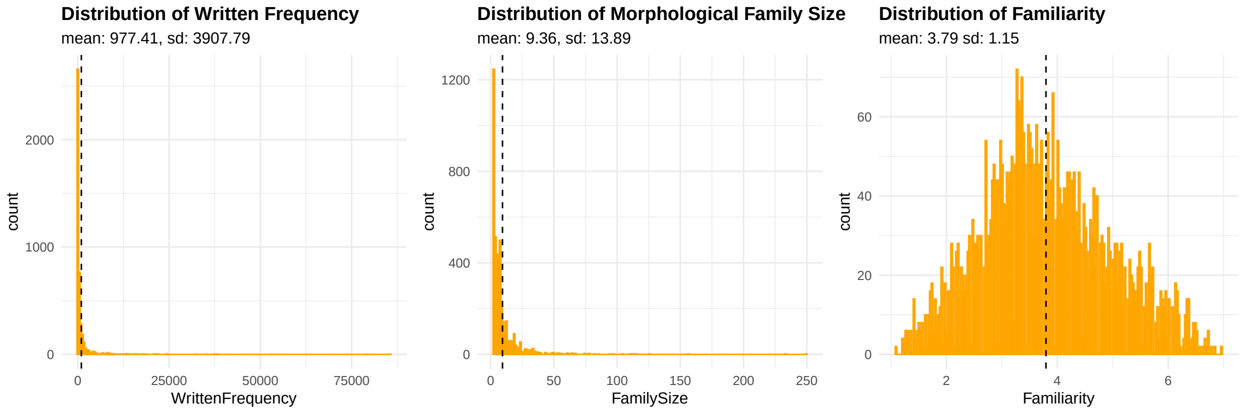
Data Exploration

Our task will revolve around identifying possible predictors for RTnaming and RTlexdec since they are the dependent variables for respective tasks. Based on our linguistic understanding, I have decided that three variables may have an important role on predicting the reaction times. They are WrittenFrequency, Familiarity and Family Size

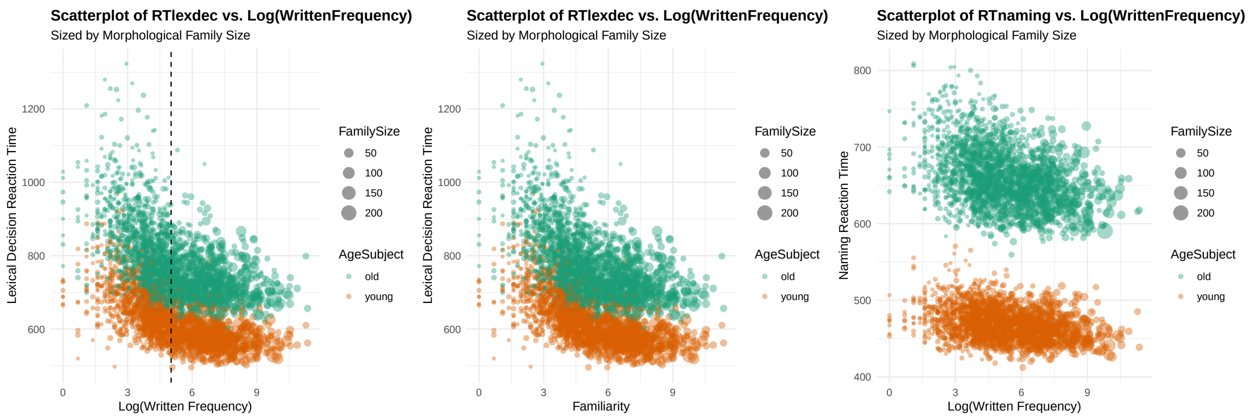
WrittenFrequency: Token frequency plays an important role in word retrieval. When a word has higher frequency, it will be more established in the memory and more easy to remember thus have a greater memory strength.

FamilySize: A Word Family refers to lexemes that are morphologically related to each other via derivation. When frequency effects apply, it can be easier to remember entries of a bigger family.

Familiarity: As a combination of different phenomena (frequency, age of acquisition, exposure) it refers to how one perceives words more familiar



We can see that frequency and family values are in zipfean distribution. Transforming these values may be necessary. For the plots below which concerns possible predictors and outcomes, I transformed them with `log()`


Scatterplot of RTlexdec vs. Log(WrittenFrequency)
Sized by Morphological Family Size

Scatterplot of RTlexdec vs. Log(WrittenFrequency)
Sized by Morphological Family Size

Scatterplot of RTnaming vs. Log(WrittenFrequency)
Sized by Morphological Family Size

Hypothesis

Given the complexity of analyzing two different two-level outcomes with three predictors; I have decided to continue with two predictors and one outcome

Null Hypothesis There is no significant relationship between lexical decision reaction times, frequency ratings and morphological family size in both age groups

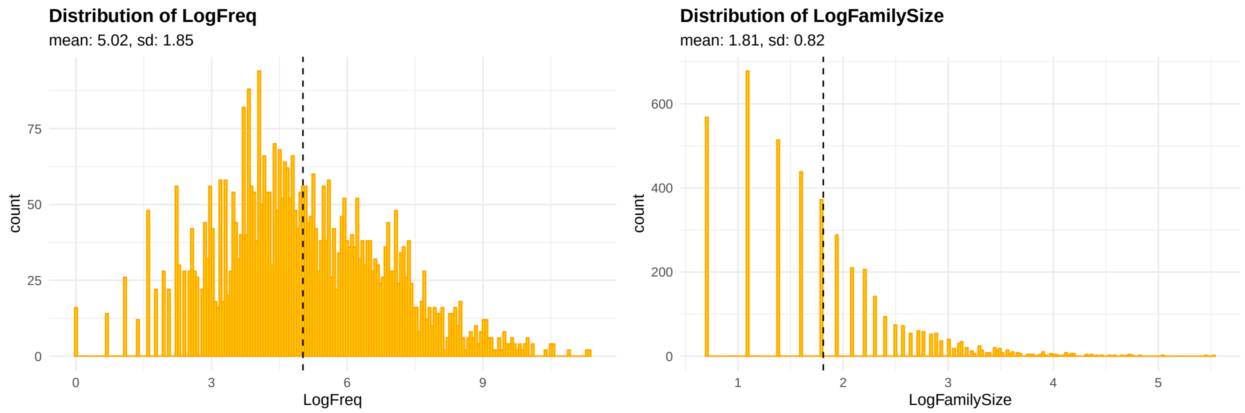
Alternative Hypothesis Lexical decision reaction times are significantly related to frequency ratings and morphological family size in both age groups.

Testing the Hypothesis

Transformations

From this point on, `eng_lexdec_naming` will be named as `df` for the ease of use. Also, unrelated columns will be dropped. I non-linear transformations:

- Log transformations of `FamilySize` and `WrittenFrequency` saved as `LogFamilySize` and `LogFreq` respectively
- AgeGroup is factorized by using `factor()`


Distribution of LogFreq
mean: 5.02, sd: 1.85

Distribution of LogFamilySize
mean: 1.81, sd: 0.82

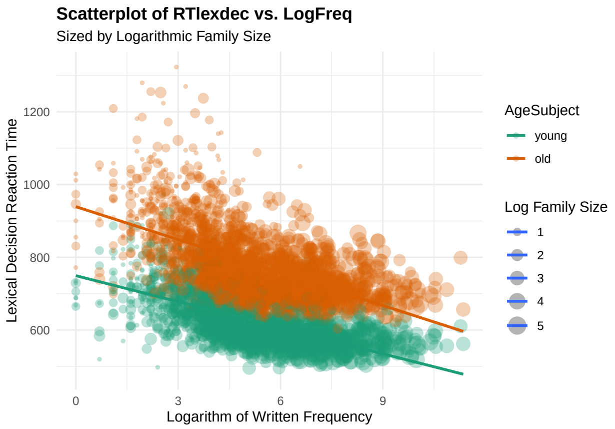
Plots show that our predictors have become more normally distributed

Models

Eight linear regression models have been built in order to test our hypothesis:

```
m1 <- lm(RTlexdec  ~  LogFreq, data=df)
m2 <- lm(RTlexdec  ~  AgeSubject, data=df)
m3 <- lm(RTlexdec  ~  LogFamilySize, data=df)
m3_2 <- lm(RTlexdec  ~  FamilySize, data=df)
m4 <- lm(RTlexdec  ~  LogFreq + AgeSubject, data=df)
m5 <- lm(RTlexdec  ~  LogFreq + LogFamilySize, data=df)
m6 <- lm(RTlexdec  ~  AgeSubject + LogFamilySize, data=df)
m7 <- lm(RTlexdec  ~  LogFreq + LogFamilySize + AgeSubject + LogFreq, data=df)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Model-1

- m1 - RTlexdec vs LogFreq
  - Our intercept is 844. It is the RT value when frequency is 0
  - Estimate of LogFreq is -27.0 . This means 1 point increase in frequency results in 27 points decrease in RT

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.188         0.188  104.     1015. 1.36e-200     1 -26642. 53291. 53310.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- p-val - Our analysis is statistically significant
- r-sq - Accounts for 18% of the variation

Model-2

- m2 - RTlexdec vs AgeSubject
    - Intercept represents the reference group “young” for AgeSubject
    - Our intercept is 630 . It is the mean RT value for young participants
    - Estimate is 158 . This means the mean of old participants is 158 point more than young participants

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.468         0.468  84.2     3866.       0     1 -25712. 51430. 51450.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- p-val - Our analysis is statistically significant
- r-sq - Accounts for 46% of the variation

Model-3

- m3 - RTlexdec vs LogFamilySize
    - Our intercept is 796 It is the RT value when LogFamilySize is 0
    - Estimate of LogFreq is -48.4 . This means 1 point increase in LogFamilySize results in 48.4 points decrease in RT

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.120         0.120  108.      598. 5.22e-124     1 -26819. 53643. 53662.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- p-val - Our analysis is statistically significant
- r-sq - Accounts for 12% of the variation

Model-3_2

- m3_2 - RTlexdec vs FamilySize
    - I wanted to try one model without transformations
    - Our intercept is 724 . It is the mean RTvalue when FamilySize is 0
    - Estimate is -1.63 . This means 1 point increase in FamilySize results in -1.63 points decrease in RT

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1    0.0384        0.0382  113.      176. 2.56e-39     1 -27013. 54032. 54051.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- p-val - Our analysis is statistically significant but decreased significantly
- r-sq - Accounts for 66% of the variation, lot lower than m3

Model-7

- m3 - RTlexdec vs LogFamilySize + AgeSubject + LogFreq
  - The most complex model
  - Intercept represents the reference group "young" for AgeSubject
  - Estimate for LogFreq and LogFamilySize is -22.7 and -14.6 respectively. Their increase similarly results in decrease of RTlexdec

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df  logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>   <dbl>  <dbl>  <dbl>
## 1     0.662         0.662  67.1     2865.       0     3 -24717. 49443. 49475.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- p-val - Our analysis is statistically significant
- r-sq - Accounts for 66% of the variation, highest of all models

---

Model Interpretation and Conclutions:

To check for which models serve the best, we check the AIC scores of the models:

```
##    df      AIC
## m7  5 49443.15
## m4  4 49519.31
## m6  4 50309.76
```

It seems that m7 is indeed our best model. Let's find out f-squared (used for effect size of multiple predictors) and conduct a power analysis.

```
## [1] "Cohen's f^2: 1.95798207989341"
```

```
## [1] "power: 1"
```

Our analysis concludes with power analysis. We have focused on analyzing `eng_lexdec_naming` dataset. After our routine exploration, we have found that three variables may play a part in predicting reaction times. For ease of analysis, we have chosen RTlexdec as outcome and FamilySize, WrittenFrequency and AgeSubject as predictors. After conducting some log transformations, seven models were built. m7 was the best model according to our metrics.

According to our analysis, our hypothesis stands. FamilySize, WrittenFrequency and AgeSubject are significant predictors when analyzing lexical decision reaction times. Old subjects react slower than young participants. Increased FamilySize and WrittenFrequency both reduced reaction times. Findings are also consistent with the literature and we have showed that results are replicable.