Initial [n] in the third person pronouns in prepositional constructions in the dialect of Mikhalevskaya village

Vasilisa Zhigulskaya

19 June 2017

Introduction

In standard Russian, all third person pronouns in prepositional constructions must have initial [n]. However, in some dialects, it is not the case. Mikhalevskaya village dialect is characterized with forms without the initial [n], which is one of its dialectal features. However, the language of the speakers is becoming standardized and there are less and less occurrences of dialectal forms. The data set used in this project consists of 1015 observations. It includes the following variables. Output variable is absence or presence of the initial [n] (categorical). Input variables are informants' ID (categorical), their year of birth (numerical), gender (categorical: male or female), education level (categorical) as well as some variables that characterize prepositional constructions: type (categorical) and frequency (numerical) of the preposition, form (categorical) and case (categorical) of the pronoun. Our hypothesis is that sociolinguistic factors (such as age of the informants, their gender and education) might influence the proportion of dialectal forms. The main idea is that the younger the speakers are, the more forms with [n] he/she has. Another (weaker) supposition is that the higher the education level (i.e. the more the contact with the standard variant), again the more cases of the initial [n] we can observe. Other variables will serve as possible predictors, although we do not know in what degree and direction they can influence the absence or presence of [n].

Description of the Phenomenon

The phenomenon illustrated in this paper is one of the linguistic variables that differs dialect of Mikhalevskaya from the standard language. It was a result of reanalysis of constructions including prepositions *vъn* 'in', *kъn* 'to' и *sъn* 'with' (later expanded to other prepositions) with third person pronouns, which took place very early in the history of Russian. In modern standard Russian, the initial nasal in pronouns is obligatory in most prepositional constructions (primary prepositions), and is optional or even impossible in constructions with some prepositions. Examples of initial [n]: *u n'ego* 'by him', *na n'ix* 'on them', *s n'im* 'with him'. On the contrary, in some Russian dialects, the initial nasal consonant after prepositions had been lost and became a dialectal feature.

Data Collection and the Data Base

The Ustja River Basin Corpus, that includes data collected in 2013 to 2016 during four field trips to Mikhalevskaya, the village in Ustya district of Arkhangelskaya Oblast, was the source of data this research is based on. It consists of of interviews, transcribed in standard Russian orthography and aligned with original audio (von Waldenfels et al. 2014). The data were collected through CQP-queries as follows:

[lemma="pronoun"] ::match.utterance_spkr="speaker". Instead of the word pronoun, a *pronoun* was included (*он*, *она* or *они*), and instead of the word *speaker*, the abbreviation of the selected speaker was included (*пфп1928*, *авм1922* etc., where part in letters is an abbreviation of speaker's name and numerical part is their year of birth). For example, the query *[lemma="он"] ::match.utterance_spkr="пфп1928"* allows us to find all forms of the pronoun *он* (singular) that were used by the speaker PFP born in 1928. The data includes third person pronouns, singular and plural, in oblique cases both in prepositional (1015 occurrences, 33 informants). Male and neuter pronouns were considered together, because they are not differentiated in the corpus annotation. Each pronominal form was examined for the presence or absence of initial nasal [n]. We did not register the initial sound in pronouns without nasal consonant (i.e. [j] or a vowel, which may itself be another parameter of variation), because there are many cases when determining the quality of the anlaut is very problematic. We only controlled whether the initial nasal [n] is present or not.

```
setwd("C:/Users/Василиса/Documents/MA_HSE/R Statistics")
df <- read.csv("pronouns.csv")
```

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

We need to check what variables might be relevant for us in the research. They are:

- `speaker` (33 informants),
- `year` (from 1922 to 1996),
- `gender` (`female` or `male`),
- `education` (`low` , `low-mid` , `high-mid` , `high`),
- `preposition` (25 prepositions),
- `prep_type` (`initial` , i.e. *в*, *к* and *с* and `later` , i.e. other prepositions)
- `case` and `form` (case and number / gender form of pronoun)
- as well as of course `consonant` (`no` and `yes`).

```
summary(df)
```

```
##     speaker         year         gender              lives
##  mdn1933: 73   Min.   :1922   female:761   Bestuzhevo   : 16
##  pfp1928: 67   1st Qu.:1933   male  :254   Mikhalevskaya:941
##  npo1965: 65   Median :1949                Plosskoe     : 58
##  avm1922: 60   Mean   :1947
##  nnt1960: 60   3rd Qu.:1960
##  lgp1947: 59   Max.   :1996
##  (Other):631
##                      born        education       index
##  Mikhalevskaya         :602   high    :352   Min.   :  1352
##  Plosskoe              :178   high-mid:282   1st Qu.:103055
##  Bestuzhevo            : 83   low     : 42   Median :204706
##  Lobanovo-Mikhalevskaya: 56   low-mid :339   Mean   :254985
##  Fomin Pochinok        : 39                  3rd Qu.:388082
##  Akichkin pochinok     : 30                  Max.   :757070
##  (Other)               : 27
##   preposition    prep_type      st_form     case     form     consonant
##  у      :581   initial:273   него   :326   acc: 55   f :257   no :560
##  с      :194   later  :742   них    :240   dat: 81   m :466   yes:455
##  к      : 74                 ней    :137   gen:649   pl:292
##  на     : 48                 ним    :123   ins:212
##  за     : 34                 нее    :120   loc: 18
##  от     : 25                 ними   : 35
##  (Other): 59                 (Other): 34
```

```
str(df)
```

```
## 'data.frame':    1015 obs. of  13 variables:
##  $ speaker    : Factor w/ 33 levels "ait1954","ans1925",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ year       : int  1954 1954 1954 1954 1954 1954 1954 1954 1954 1954 ...
##  $ gender     : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lives      : Factor w/ 3 levels "Bestuzhevo","Mikhalevskaya",..: 2 2 2 2 2 2
2 2 2 2 ...
##  $ born       : Factor w/ 8 levels "Akichkin pochinok",..: 6 6 6 6 6 6 6 6 6 6
...
##  $ education  : Factor w/ 4 levels "high","high-mid",..: 1 1 1 1 1 1 1 1 1 1 ..
.
##  $ index      : int  197434 197446 216169 425772 228784 230299 230310 425760 67
2850 673064 ...
##  $ preposition: Factor w/ 25 levels "без","в","для",..: 24 24 24 24 23 8 24 24
24 24 ...
##  $ prep_type  : Factor w/ 2 levels "initial","later": 2 2 2 2 1 1 2 2 2 2 ...
##  $ st_form    : Factor w/ 8 levels "него","нее","ней",..: 8 1 1 1 7 6 8 8 1 1 .
..
##  $ case       : Factor w/ 5 levels "acc","dat","gen",..: 3 3 3 3 4 2 3 3 3 3 ..
.
##  $ form       : Factor w/ 3 levels "f","m","pl": 3 2 2 2 3 3 3 3 2 2 ...
##  $ consonant  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 2 2 ...
```

In order to see whether the frequency of the preposition has any influence on the pronunciation of the following pronoun, we need to calculate and add this information. For that, we create a separate table with frequencies and then add them in a column to the data frame with the help of `inner_join` (it is a numerical variable):

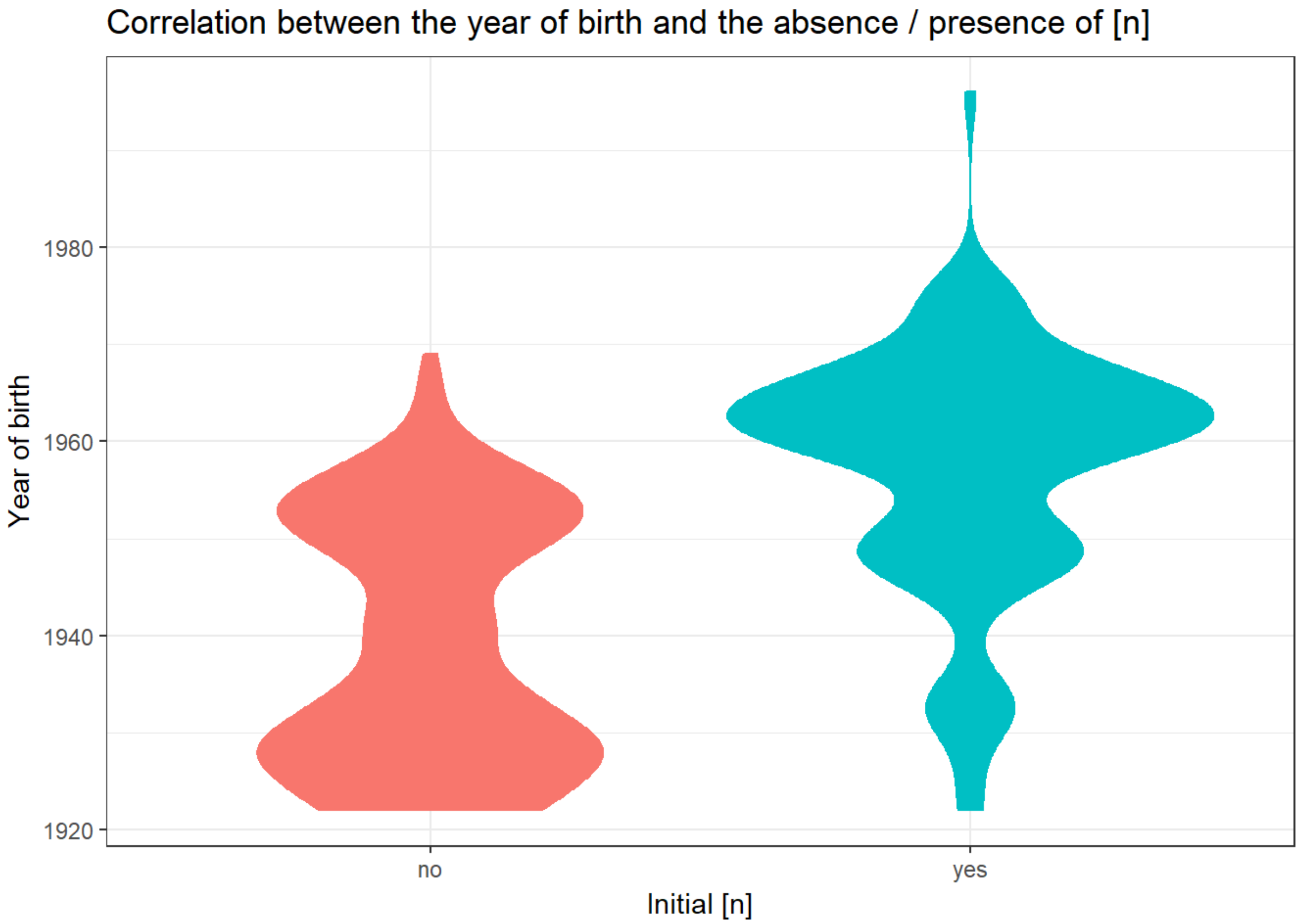
```
df %>%
  group_by(preposition) %>%
  summarise(prep_frequency = n()/1015) ->
  df_freq
df <- inner_join(df, df_freq)
```

```
## Joining, by = "preposition"
```

Descriptive Statistics

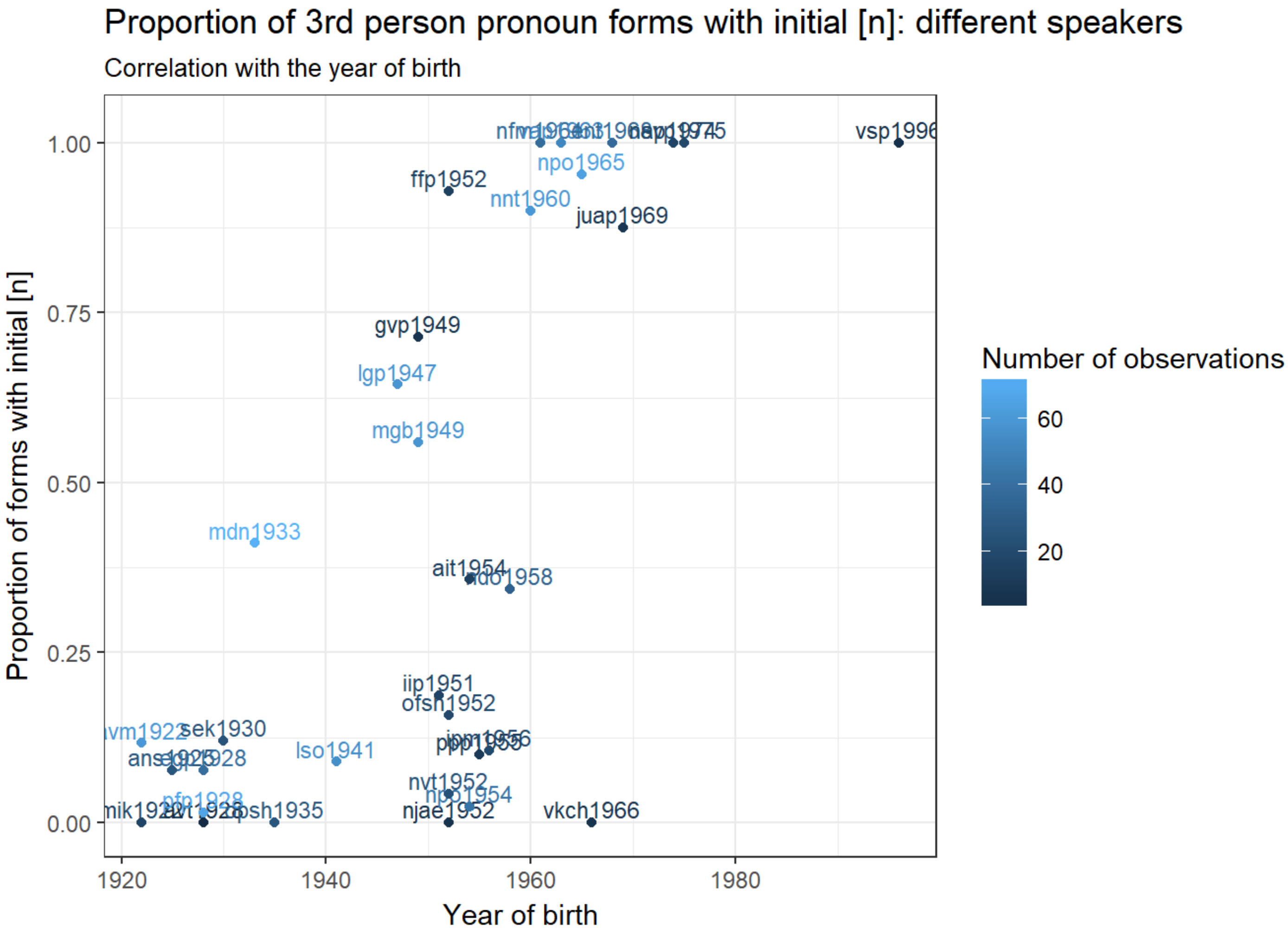
To begin with, we want to visualize the correlation between the year of birth and the absence or presence of [n]. For first, we will not differentiate between the speakers in order to see the general tendency. We draw a violin plot and see that, in general, there is a trend to have more observations without [n] among older speakers and with [n] among younger ones. But we must be careful because this kind of visualization does not take into account how many utterances alltogether there are in the interview from one speaker. It might display not the tendency but the disproportionality of the collected data.

```
df %>%
  ggplot(aes(consonant, year, fill = consonant, color = consonant)) +
  geom_violin(show.legend = FALSE) +
  labs(title = "Correlation between the year of birth and the absence / presence o
f [n]", x = "Initial [n]", y = "Year of birth") +
  theme_bw()
```



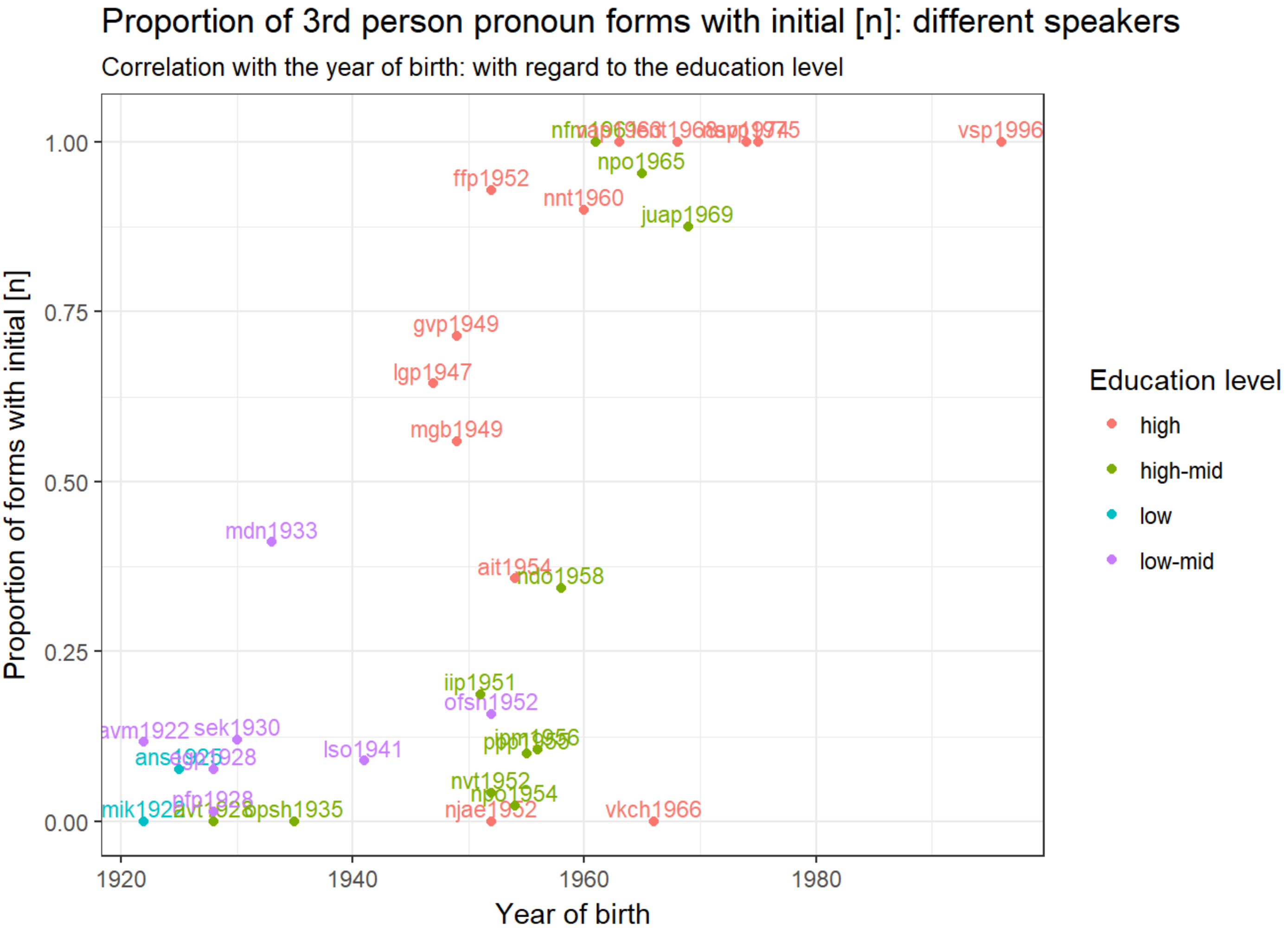
Therefore, let us draw a scatter plot considering each speaker separately in order to check the correlation between the year of birth and the empirical proportion of observations with [n]. As these are only observed proportions and not absolute measures, we also want to see and keep in mind what is the number of observations.

```
df %>%
  group_by(year, speaker, education) %>%
  summarise(prop_consonant = sum(consonant == "yes")/(sum(consonant == "yes") + su
m(consonant == "no")), all_consonant = (sum(consonant == "yes") + sum(consonant ==
"no"))) %>%
  ggplot(aes(year, prop_consonant, color = all_consonant, label = speaker)) +
  geom_text(nudge_y = 0.02, size = 3) +
  geom_point() +
  labs(title = "Proportion of 3rd person pronoun forms with initial [n]: different
speakers", subtitle = "Correlation with the year of birth", x = "Year of birth", y
= "Proportion of forms with initial [n]", color = "Number of observations") +
  theme_bw()
```



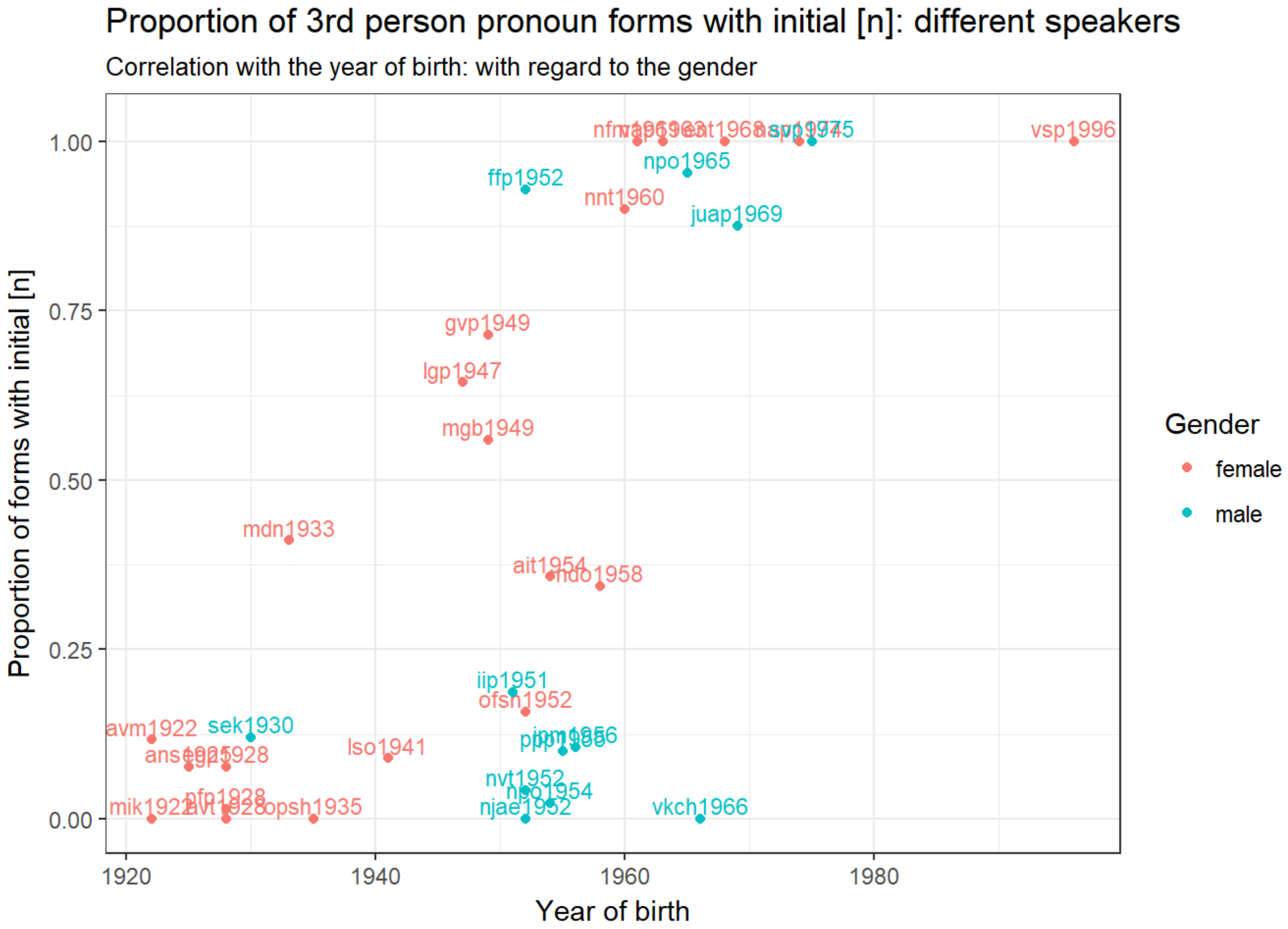
As we also suppose that education level might have an impact on the dialectal performance on the speakers, we need to visualize it first. Let us display the education level on our scatter plot. We can observe that it is probably not fully independent variable and depends on the year of birth. Therefore, in our analysis we should keep in mind the option to consider the integration of these variables.

```
df %>%
  group_by(year, speaker, education) %>%
  summarise(prop_consonant = sum(consonant == "yes")/(sum(consonant == "yes") + su
m(consonant == "no"))) %>%
  ggplot(aes(year, prop_consonant, colour = education, label = speaker)) +
  geom_text(nudge_y = 0.02, size = 3, show.legend = FALSE) +
  geom_point() +
  labs(title = "Proportion of 3rd person pronoun forms with initial [n]: different
speakers", subtitle = "Correlation with the year of birth: with regard to the educ
ation level", x = "Year of birth", y = "Proportion of forms with initial [n]", col
or = "Education level") +
  theme_bw()
```



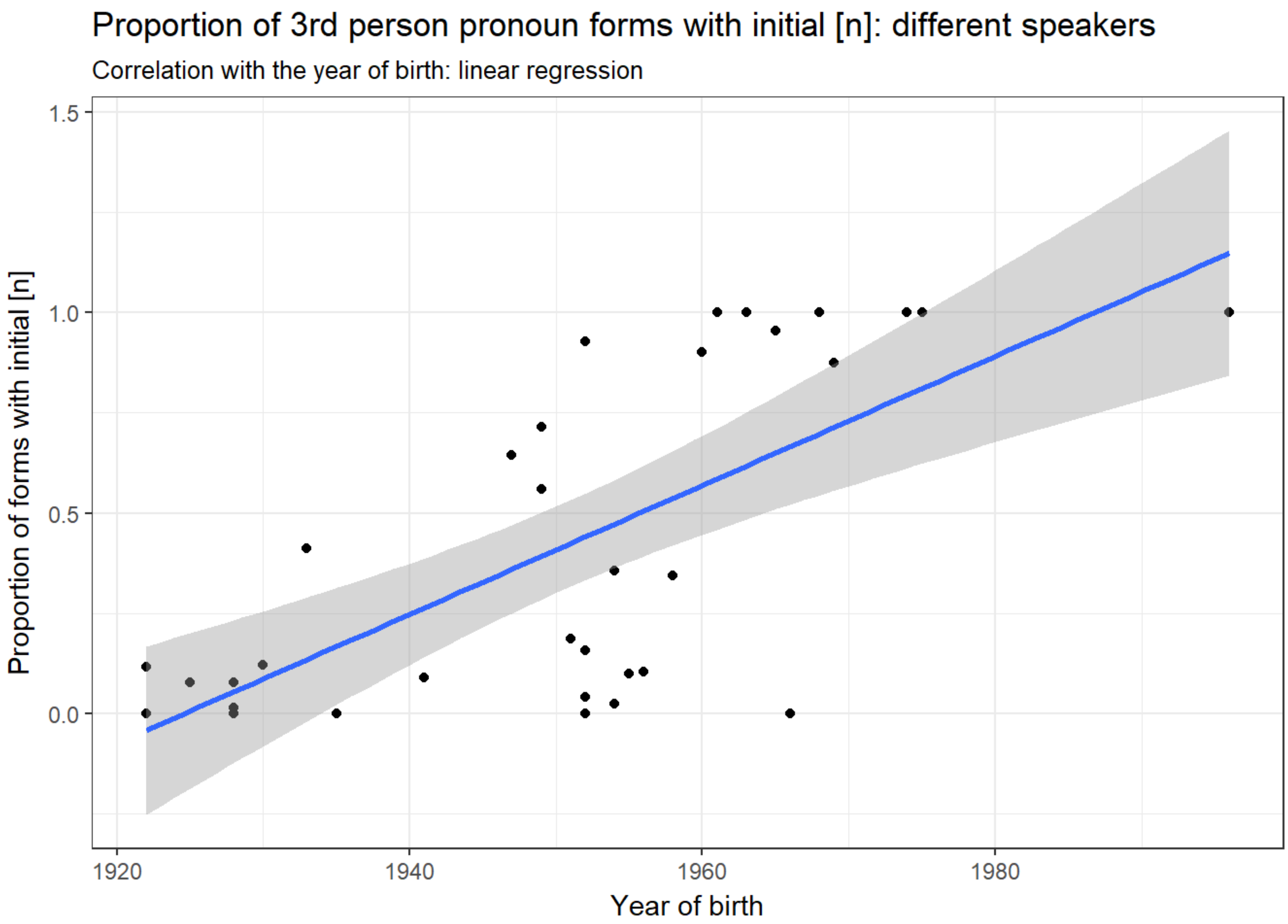
We also suppose the dependency on the gender, so let us display this variable on the scatter plot. Again, male speakers are mostly born in 1950-1970, so probably the sample is not perfect for the analysis and depends on the year. We chould check the correlation with the statistical methods.

```
df %>%
  group_by(year, speaker, gender) %>%
  summarise(prop_consonant = sum(consonant == "yes")/(sum(consonant == "yes") + su
m(consonant == "no"))) %>%
  ggplot(aes(year, prop_consonant, color = gender, label = speaker)) +
  geom_text(nudge_y = 0.02, size = 3, show.legend = FALSE) +
  geom_point() +
  labs(title = "Proportion of 3rd person pronoun forms with initial [n]: different
speakers", subtitle = "Correlation with the year of birth: with regard to the gend
er", x = "Year of birth", y = "Proportion of forms with initial [n]", color = "Gen
der") +
  theme_bw()
```



First, we want to check whether the linear regression model is good for our data. In order to do that, we should transform our data frame into a shorter format, so that each observation is not a pronoun with preposition but a speaker with a certain number or dialectal `dial` (without [n]) and innovative `inn` (with [n]) pronunciations. Then we plot our linear regression with the predictor `year` .

```
df %>%
  group_by(speaker, year, gender, education) %>%
  summarise(dial = sum(consonant=="no"), inn = sum(consonant=="yes")) ->
  num_df
num_df %>%
  mutate(perc = inn/(dial + inn)) %>%
  ggplot(aes(year, perc))+
  geom_point()+
  geom_smooth(method = "lm") +
  labs(title = "Proportion of 3rd person pronoun forms with initial [n]: different
speakers", subtitle = "Correlation with the year of birth: linear regression", x =
"Year of birth", y = "Proportion of forms with initial [n]") +
  theme_bw()
```



After plotting the linear regression we see two major problems: 1. It is not good, because it does not cover all the variability (or the big part of it); 2. It predicts values above 1 and less than 0, which is impossible, because a speaker cannot pronounce less than 0 per cent and more then 100 per cent. Therefore, we need to consider our data as the one with binary dependent variable. Only in this case the model will be able to give us realistic results.

Logistic Regression Model

The data set analyzed in our research consists of binary variables, where only two outcomes are possible: variable `consonant` with the value `no` (the absence of initial [n]) and with the value `yes` (the presence of initial [n]). Therefore, the calculations were made with the help of Logit. The Logistic Regression Model has advantages for our data. First, predicted values are always between 0 and 1. It means that predicted proportions won't be above 1 or below 0, which is impossible in our context. Second, in comparison to simple proportions, this method allows to weight the amount of contributions in each case (the number of occurrences for every informant). This is very important in case of data like ours, because some speakers have extremely low number of outcomes. In R, logits can be easily calculated with the function `glm()` (i.e. Generalized Linear Model). The hypotheses that we are going to check with the help of logistic regression model are:

- H0: There is no correlation between `consonant` and the predictors (independent variables and their interactions).
- H1: There is a correlation between `consonant` and the predictors (independent variables and their interactions).

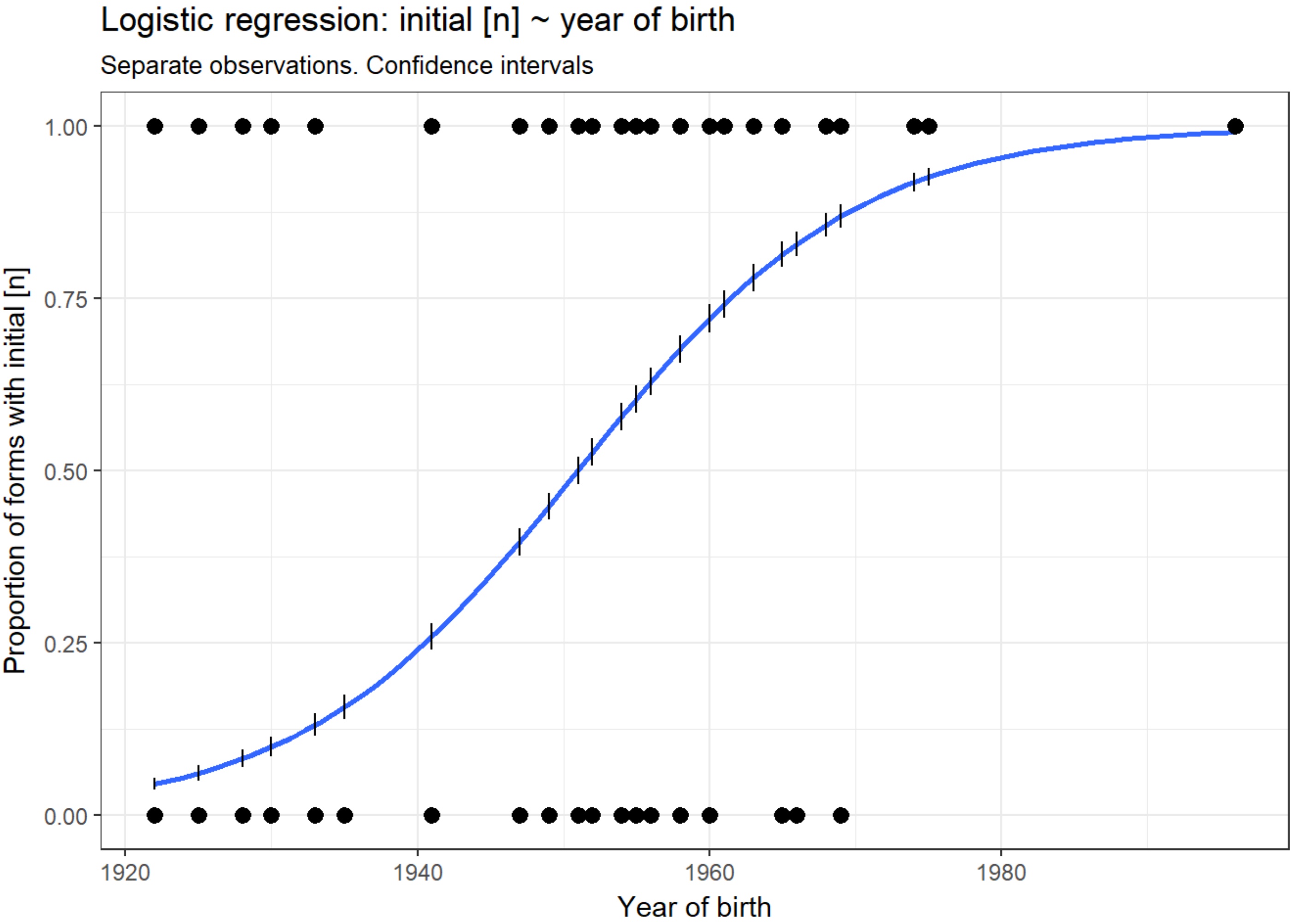
First, let us consider a simple model with the numerical variable `year` as a predictor. Then we can visulize it and compare with the previuos results.

```
fit_year <- glm(consonant~year, data = df, family = "binomial")
summary(fit_year)
```

```
##
## Call:
## glm(formula = consonant ~ year, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0174  -0.7745  -0.3044   0.7735   2.4881
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.052e+02  1.295e+01  -15.85   <2e-16 ***
## year         1.052e-01  6.636e-03   15.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1396.21  on 1014  degrees of freedom
## Residual deviance:  988.72  on 1013  degrees of freedom
## AIC: 992.72
##
## Number of Fisher Scoring iterations: 5
```

As a result, the estimate coefficient, that shows dependence on predictor `year` , equals to `1.052e-01` , and is positive, which means that the later the informants were born the higher is the probability that they give innovative responses with [n]. The significance code mentioned in the model is `***` , the dependence is considered to be significant, i.e. the coefficient rate cannot be explained by randomness. Let us plot the sigmoid for this logistic regression, together with confidence intervals.

```
df_ci <- cbind.data.frame(df, predict(fit_year, df, type = "response", se.fit = TRUE)[1:2])
df_ci %>%
  mutate(`P(consonant)` = as.numeric(consonant) - 1) %>%
  ggplot(aes(x = year, y = `P(consonant)`))+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  geom_point() +
  geom_pointrange(aes(x = year, ymin = fit - se.fit, ymax = fit + se.fit))+
  labs(title = "Logistic regression: initial [n] ~ year of birth", subtitle = "Separate observations. Confidence intervals", x = "Year of birth", y = "Proportion of forms with initial [n]") +
  theme_bw()
```

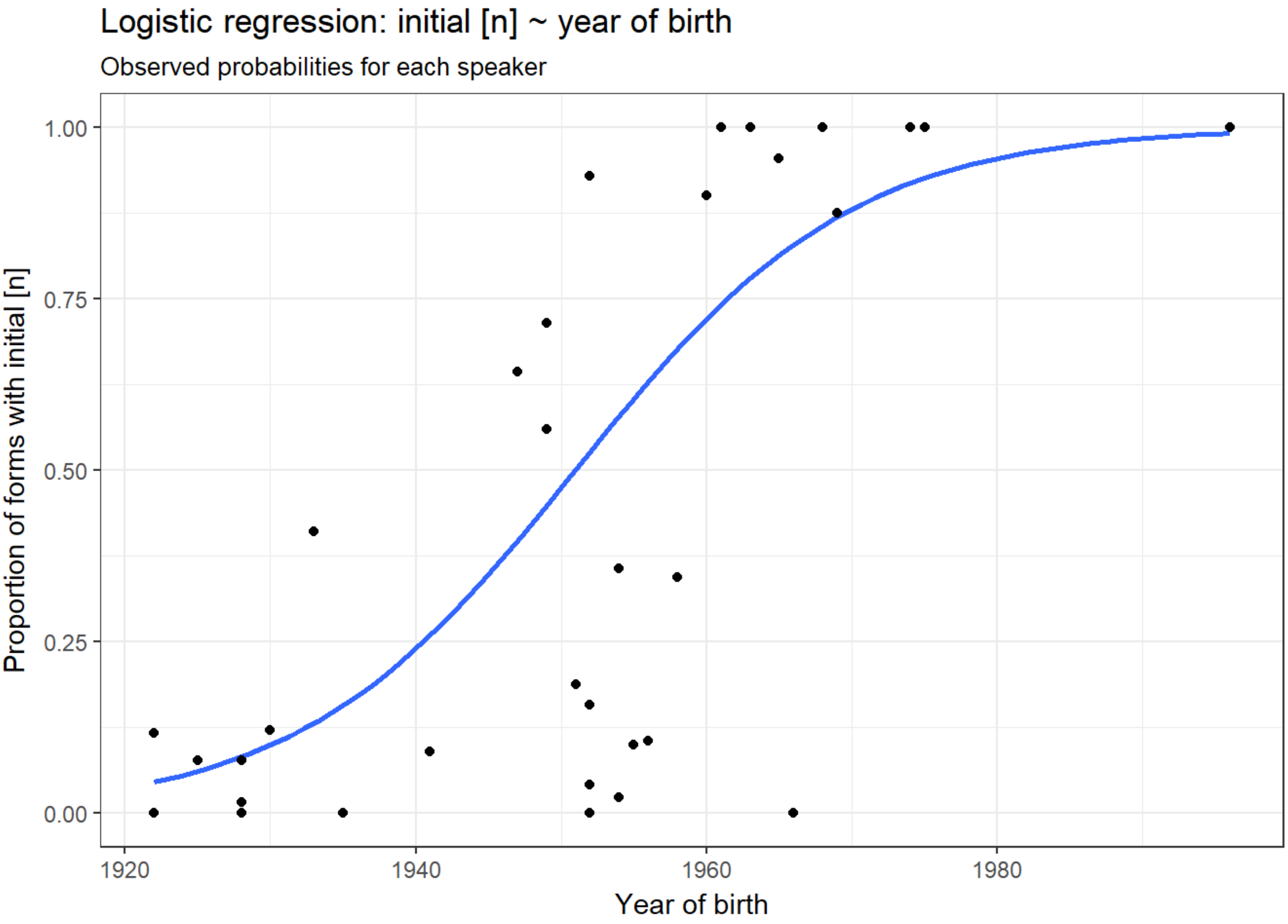


The plot in shape of S curve predicts the distribution of probability of variable value [n] among speakers of different year of birth. Unfortunately, this plot does not tell us much about how well this sigmoid displays the actual perdormance of each speaker. Moreover, confidence intervals do not provide us with much information. We need to plot the observed probabilities besides the sigmoid with the predicted ones (first, create a data frame with probabilities `df_probs` , then join it with the regular data frame and then plot separate points for the observed probabilities).

```
df %>%
  group_by(year, speaker, education) %>%
  summarise(prop_consonant = sum(consonant == "yes")/(sum(consonant == "yes") + su
m(consonant == "no"))) ->
  df_probs
df <- inner_join(df, df_probs)
```

```
## Joining, by = c("speaker", "year", "education")
```

```
df %>%
  mutate(`P(consonant)` = as.numeric(consonant) - 1) %>%
  ggplot(aes(x = year, y = `P(consonant)`))+
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE)
+
  geom_point(aes(x = year, y = prop_consonant)) +
  labs(title = "Logistic regression: initial [n] ~ year of birth", subtitle = "Obs
erved probabilities for each speaker", x = "Year of birth", y = "Proportion of for
ms with initial [n]") +
  theme_bw()
```



After we drew a plot for the numerical predictor, let us see what are the predictions of the model with all possible predictors:

```
fit_all <- glm(consonant ~ year + education + gender + prep_type + prep_frequency
+ case + form, data = df, family = "binomial")
summary(fit_all)
```

```
##
## Call:
## glm(formula = consonant ~ year + education + gender + prep_type +
##     prep_frequency + case + form, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3726  -0.6714  -0.2535   0.7094   2.8196
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -273.45609   24.15450 -11.321  < 2e-16 ***
## year                 0.14078    0.01237  11.379  < 2e-16 ***
## educationhigh-mid   -1.23655    0.25070  -4.932 8.12e-07 ***
## educationlow        -0.08765    0.82183  -0.107 0.915061
## educationlow-mid    -0.12745    0.30256  -0.421 0.673583
## gendermale          -0.82000    0.24187  -3.390 0.000698 ***
## prep_typelater      -0.48163    0.44711  -1.077 0.281384
## prep_frequency       0.82877    0.61910   1.339 0.180679
## casedat              0.21329    0.62320   0.342 0.732160
## casegen              0.16175    0.49409   0.327 0.743383
## caseins             -0.29036    0.57535  -0.505 0.613791
## caseloc              0.36159    0.88075   0.411 0.681404
## formm               -0.82600    0.21877  -3.776 0.000160 ***
## formpl              -0.29916    0.24332  -1.230 0.218876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1396.21  on 1014  degrees of freedom
## Residual deviance:  862.35  on 1001  degrees of freedom
## AIC: 890.35
##
## Number of Fisher Scoring iterations: 5
```

The first thing that we can observe is that predictors `prep_type` , `case` and `prep_frequency` are not significant for out model (according to the significance code). But male `form` of the pronoun changes the log odds of initial [n] by `-0.82600` . Both `year` and `gender` are statistically significant, as is one term for `education` (`educationhigh-mid`):

- For every one unit change in `year` , the log odds of initial [n] (versus its absence) increases by `0.14078` .
- Being a `male` versus `female` changes the log odds of initial [n] by `-0.82000` .
- Having the `education` of the level `high-mid` versus the level `high` , changes the log odds of initial [n] by `-1.23655` .

After applying the Anova test, we also see that the strongest predictor is the year of birth `year` : it allows to get rid of the biggest amount of deviance. On the contrary, `prep_type` , `case` and `prep_frequency` are almost unsignificant. This means that the preposition probably has no influence on the absence or presence of the initial [n].

```
anova(fit_all, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: consonant
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             1014    1396.21
## year             1   407.49      1013     988.72 < 2.2e-16 ***
## education        3    86.37      1010     902.35 < 2.2e-16 ***
## gender           1    15.31      1009     887.04  9.12e-05 ***
## prep_type        1     0.00      1008     887.04 0.9788970
## prep_frequency   1     4.88      1007     882.17 0.0272365 *
## case             4     3.52      1003     878.65 0.4748574
## form             2    16.30      1001     862.35 0.0002891 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, we can remove predictors `prep_type` , `case` and `prep_frequency` from our model. Model `fit4` is the model only with significant predictors. Its `AIC` is `885.59` .

```
fit4 <- glm(consonant ~ year + education + gender + form, data = df, family = "bin
omial")
summary(fit4)
```

```
##
## Call:
## glm(formula = consonant ~ year + education + gender + form, family = "binomial"
,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2976  -0.7046  -0.2743   0.7131   2.7668
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -272.65879   24.08275 -11.322  < 2e-16 ***
## year                 0.14040    0.01234  11.377  < 2e-16 ***
## educationhigh-mid   -1.30552    0.24841  -5.255 1.48e-07 ***
## educationlow        -0.09589    0.81898  -0.117 0.906795
## educationlow-mid    -0.15851    0.30041  -0.528 0.597749
## gendermale          -0.80681    0.23959  -3.367 0.000759 ***
## formm               -0.84165    0.21470  -3.920 8.85e-05 ***
## formpl              -0.29707    0.23553  -1.261 0.207191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1396.21  on 1014  degrees of freedom
## Residual deviance:  869.59  on 1007  degrees of freedom
## AIC: 885.59
##
## Number of Fisher Scoring iterations: 5
```

As a next step we need to create a table to look and the predicted probabilities of this model for different values of independent variables.

```
df %>%
  count(year, education, gender, form, consonant) %>%
  select(-n, -consonant) %>%
  unique() ->
  fit_prob_df
fit_prob_df %>%
  predict(fit4, newdata = ., type = "response") ->
  fit_prob_df$prediction
fit_prob_df %>%
  arrange(prediction)
```

```
## # A tibble: 86 x 5
##     year education gender   form prediction
##    <int>    <fctr> <fctr> <fctr>      <dbl>
##  1  1928  high-mid female      m 0.01613983
##  2  1922   low-mid female      m 0.02176185
##  3  1922       low female      m 0.02313566
##  4  1930   low-mid   male      m 0.02962140
##  5  1925       low female      m 0.03483201
##  6  1922   low-mid female     pl 0.03693281
##  7  1922       low female     pl 0.03922595
##  8  1935  high-mid female      m 0.04199200
##  9  1922   low-mid female      f 0.04908140
## 10  1928   low-mid female      m 0.04911725
## # ... with 76 more rows
```

```
fit_prob_df %>%
  arrange(desc(prediction))
```

```
## # A tibble: 86 x 5
##     year education gender   form prediction
##    <int>    <fctr> <fctr> <fctr>      <dbl>
##  1  1996      high female      f  0.9994919
##  2  1996      high female     pl  0.9993163
##  3  1974      high female      f  0.9889639
##  4  1974      high female     pl  0.9852028
##  5  1975      high   male      f  0.9787324
##  6  1974      high female      m  0.9747619
##  7  1968      high female      f  0.9747430
##  8  1975      high   male     pl  0.9715847
##  9  1968      high female     pl  0.9663006
## 10  1975      high   male      m  0.9520029
## # ... with 76 more rows
```

As we see, the lowest probability of initial [n] `0.01613983` is for the male pronoun in the speech of female informant 1928 year of birth and high-mid education level. The highest `0.9994919` for the female pronoun in the speech of female informant 1996 year of birth and high education level. This data is hardly interpetable because we mix socolinguitic predictors and observation-characteristic predictor. For this table, let us consider solely sociolinguitic factors (to understand something about the speakers of the dialect).

```
df %>%
  count(year, education, gender, consonant) %>%
  select(-n, -consonant) %>%
  unique() ->
  proba_df
fit_socio <- glm(consonant ~ year + education + gender, data = df, family = "binom
ial")
proba_df %>%
  predict(fit_socio, newdata = ., type = "response") ->
  proba_df$prediction
proba_df %>%
  arrange(prediction)
```

```
## # A tibble: 30 x 4
##     year education gender prediction
##    <int>    <fctr> <fctr>      <dbl>
##  1  1928  high-mid female 0.02796769
##  2  1922   low-mid female 0.03506336
##  3  1922       low female 0.03685432
##  4  1930   low-mid   male 0.04137803
##  5  1925       low female 0.05424350
##  6  1935  high-mid female 0.06888340
##  7  1928   low-mid female 0.07547728
##  8  1933   low-mid female 0.13813156
##  9  1951  high-mid   male 0.20545696
## 10  1952  high-mid   male 0.22835545
## # ... with 20 more rows
```

```
proba_df %>%
  arrange(desc(prediction))
```

```
## # A tibble: 30 x 4
##     year education gender prediction
##    <int>    <fctr> <fctr>      <dbl>
##  1  1996      high female  0.9989835
##  2  1974      high female  0.9805907
##  3  1975      high   male  0.9589163
##  4  1968      high female  0.9574235
##  5  1963      high female  0.9197085
##  6  1960      high female  0.8842864
##  7  1966      high   male  0.8739123
##  8  1954      high female  0.7728024
##  9  1969  high-mid   male  0.7457076
## 10  1961  high-mid female  0.7117072
## # ... with 20 more rows
```

This provides us only with the important information . Now we can say what are the characteristics of the most and the least dialectal informant (not depending on the pronoun): the lowest probability of [n] outcome `0.02796769` is for the the female speaker 1928 year of birth and high-mid education level and the highest `0.9989835` for the the female speaker 1996 year of birth and high education level (the same as in th previous table).

However, before removing the variable `case` , let us try to look at its integration with the variable `form` because they both constitute the output form of a pronoun (i.e. case and gender). The problem with this model is that it is too complicated.

```
fit5 <- glm(consonant ~ year + education + gender + case*form, data = df, family =
"binomial")
anova(fit4, fit5)
```

```
## Analysis of Deviance Table
##
## Model 1: consonant ~ year + education + gender + form
## Model 2: consonant ~ year + education + gender + case * form
##   Resid. Df Resid. Dev Df Deviance
## 1      1007     869.59
## 2       995     848.59 12   20.996
```

We see that we probably should not fully remove the predictor `case` but consider it in the integration with `form` . Moreover, we want to check the integration of `year` , `education` and `gender` , because they constitute the group of sociolinguistic factors and may influence each other.

```
fit6 <- glm(consonant ~ year * education + gender + case*form, data = df, family =
"binomial")
fit_final <- glm(consonant ~ year * education * gender + case*form, data = df, fam
ily = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
anova(fit5, fit6)
```

```
## Analysis of Deviance Table
##
## Model 1: consonant ~ year + education + gender + case * form
## Model 2: consonant ~ year * education + gender + case * form
##   Resid. Df Resid. Dev Df Deviance
## 1       995     848.59
## 2       992     725.82  3   122.78
```

```
anova(fit6, fit_final)
```

```
## Analysis of Deviance Table
## 
## Model 1: consonant ~ year * education + gender + case * form
## Model 2: consonant ~ year * education * gender + case * form
##   Resid. Df Resid. Dev Df Deviance
## 1       992     725.82            
## 2       988     699.33  4   26.484
```

We can observe that the model with integration of all sociolinguistic factors covers more variability than the simple additive model. The AIC (`753.33`) is the lowest for this model in comparison to the other presented fits. This is going to be our final model (`fit_final`).

```
summary(fit_final)
```

```
## 
## Call:
## glm(formula = consonant ~ year * education * gender + case * 
##     form, family = "binomial", data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.1614  -0.5406  -0.1675   0.3920   2.8842  
## 
## Coefficients: (3 not defined because of singularities)
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -3.662e+02  5.727e+01  -6.394 1.62e-10
## year                               1.875e-01  2.936e-02   6.389 1.67e-10
## educationhigh-mid                 -1.164e+04  4.279e+05  -0.027  0.97831
## educationlow                      -1.114e+04  2.810e+06  -0.004  0.99684
## educationlow-mid                   3.059e+02  6.845e+01   4.468 7.88e-06
## gendermale                         2.348e+02  9.070e+01   2.589  0.00963
## casedat                            2.700e+00  1.202e+00   2.247  0.02466
## casegen                            1.853e+00  1.134e+00   1.634  0.10227
## caseins                            1.430e+00  1.165e+00   1.228  0.21959
## caseloc                            2.131e+00  1.430e+00   1.490  0.13614
## formm                              1.329e-01  1.278e+00   0.104  0.91721
## formpl                             1.306e+00  1.351e+00   0.967  0.33340
## year:educationhigh-mid             5.942e+00  2.186e+02   0.027  0.97831
## year:educationlow                  5.789e+00  1.460e+03   0.004  0.99684
## year:educationlow-mid             -1.579e-01  3.519e-02  -4.487 7.21e-06
## year:gendermale                   -1.206e-01  4.636e-02  -2.602  0.00927
## educationhigh-mid:gendermale       1.083e+04  4.279e+05   0.025  0.97980
## educationlow:gendermale                   NA         NA      NA       NA
## educationlow-mid:gendermale       -2.018e+00  1.462e+00  -1.380  0.16755
## casedat:formm                     -2.395e+00  1.541e+00  -1.554  0.12008
## casegen:formm                     -1.283e+00  1.326e+00  -0.967  0.33332
## caseins:formm                     -6.536e-01  1.361e+00  -0.480  0.63115
## caseloc:formm                     -1.943e+00  2.497e+00  -0.778  0.43642
## casedat:formpl                    -4.658e+00  1.651e+00  -2.821  0.00478
## casegen:formpl                    -1.692e+00  1.389e+00  -1.219  0.22296
## caseins:formpl                    -2.023e+00  1.485e+00  -1.362  0.17309
## caseloc:formpl                    -4.612e+00  2.243e+00  -2.056  0.03976
## year:educationhigh-mid:gendermale -5.532e+00  2.186e+02  -0.025  0.97981
## year:educationlow:gendermale              NA         NA      NA       NA
## year:educationlow-mid:gendermale          NA         NA      NA       NA
##
## (Intercept)                       ***
## year                              ***
## educationhigh-mid
## educationlow
## educationlow-mid                  ***
## gendermale                        **
## casedat                           *
## casegen
## caseins
## caseloc
## formm
## formpl
## year:educationhigh-mid
## year:educationlow
## year:educationlow-mid             ***
## year:gendermale                   **
## educationhigh-mid:gendermale
## educationlow:gendermale
## educationlow-mid:gendermale
## casedat:formm
## casegen:formm
## caseins:formm
## caseloc:formm
## casedat:formpl                    **
## casegen:formpl
## caseins:formpl
## caseloc:formpl                    *
## year:educationhigh-mid:gendermale
## year:educationlow:gendermale
## year:educationlow-mid:gendermale
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1396.21  on 1014  degrees of freedom
## Residual deviance:  699.33  on  988  degrees of freedom
## AIC: 753.33
##
## Number of Fisher Scoring iterations: 19
```

Summary

Even on our simple descriptive scatter plots we could see that there might be some correlation between the proportion of the observations with the initial [n] in the forms of third person pronouns in prepositional constructions and year. We also could suppose that education level and gender may influence this variable. We showed that linear regression model does not fit to the data like this: altough all speakers have their own degree of being dialectal, we must keep considering the data as having binary dependent variable. After trying different possible models we came to the following results. The simplest and, nonetheless, quite strong, model is with the sole predictor `year` year of birth (it explains the most number of variability). We also can add other predictors, which, however, complicate the model and each of them is less significant. One more important result is the observation that, apparently, the preposition (its type and frequency) has no impact on the choice of the form.

Sources

- Baayen R.H. (2008). *Analyzing Linguistic Data: a Practical Introduction to Statistics Using R*. Cambridge University Press.
- Daniel M., Dobrushina N., von Waldenfels R.. *The language of the Ustja river bassin. A corpus of North Russian dialectal speech*. 2013-2014. Bern, Moscow. Electronic resource.
- Gries, St. Th. (2013). *Statistics for Linguistics with R: A Practical Introduction*. 2nd ed. Mouton de Gruyter.
- von Waldenfels R., Daniel M., Dobrushina N. (2014). *Why Standard Orthography? Building the Ustya River Basin Corpus, an online corpus of a Russian dialect*. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2014” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy Konferentsii «Dialog 2014»], Bekasovo, Vyp. 13 (20).
- Zhigulskaya V.R. (2015). *A variation study of third person pronouns in prepositional constructions: a North Russian Dialect*. Paper presented at the The International Conference on Language Variation in Europe (ICLaVE) 8, Leipzig, Germany, 2015, May.