Data, data documentation and analysis scripts for

Determinants of English accents.

Martijn Wieling[(1)] & Jelke Bloem[(2)] & R. Harald Baayen[(4,5)] & John Nerbonne[(1,6)]

[1]University of Groningen, the Netherlands & [2]University of Amsterdam, the Netherlands & [3]Utrecht University, the Netherlands & [4]Eberhard Karls University, Germany & [5]University of Alberta, Canada & [6]University of Freiburg, Germany

Status: Submitted
Preprint: `http://www.martijnwieling.nl/files/WielingEtAl-accents-submitted.pdf`

Abstract

In this study we investigate which factors affect the degree of non-native accent of L2 speakers of English who learned English in school and mostly lived for some time in an anglophone setting. We use data from the Speech Accent Archive containing over 700 speakers speaking almost 160 different native languages. We show that besides several important predictors, including the age of English onset and length of anglophone residence, the linguistic distance between the speaker's native language and English is a significant predictor of the degree of non-native accent in pronunciation. This study extends an earlier study (Schepens et al., 2013) which only focused on Indo-European L2 learners of Dutch and used a general speaking proficiency measure.
Keywords: Second language learning, Mixed-effects regression, Pronunciation nativelikeness, Linguistic distances.

1 Packages and functions

```
library(mgcv)
library(itsadug)
library(car)

R.Version()$version.string

## [1] "R version 3.2.1 (2015-06-18)"

packageVersion('mgcv')

## [1] '1.8.6'

packageVersion('itsadug')

## [1] '1.0.1'

packageVersion('car')

## [1] '2.0.25'
```

2 English accents data set

```
load("data/accents.rda")
```

Legenda `accents` (712 observations of 22 variables):

1. Speaker : the speaker
2. Language : the native language of the speaker
3. Iso : iso code of the language (as used in ASJP)
4. Country : the country of birth of the speaker
5. LD.log : log-transformed average Levenshtein distance with respect to average native American English (see Wieling, Bloem, Mignella et al., 2014, *Language Dynamics and Change*)
6. Nativelikeness : the human-rated nativelikeness of the speaker (only available for 237 speakers; also from Wieling, Bloem, Mignella et al., 2014)
7. IsIESpeaker : if the speaker has an Indo-European Language (TRUE) or not (FALSE)
8. Age : the age of the speaker
9. IsMale : the gender of the speaker (1: male, 0: female)
10. AEO : the age of English onset of the speaker
11. LR : the cumulative length of residence in an English-speaking country of a speaker
12. NrLang : the number of additional languages the speaker speaks (besides English)
13. GNI.log : the log-transformed gross national income of the speaker's country of birth (in 2011)
14. CountryEduYrs : the average number of years of education in the speaker's country (in 2011)
15. *.c: centered numerical predictors

3 Analysis and results

3.1 Descriptives

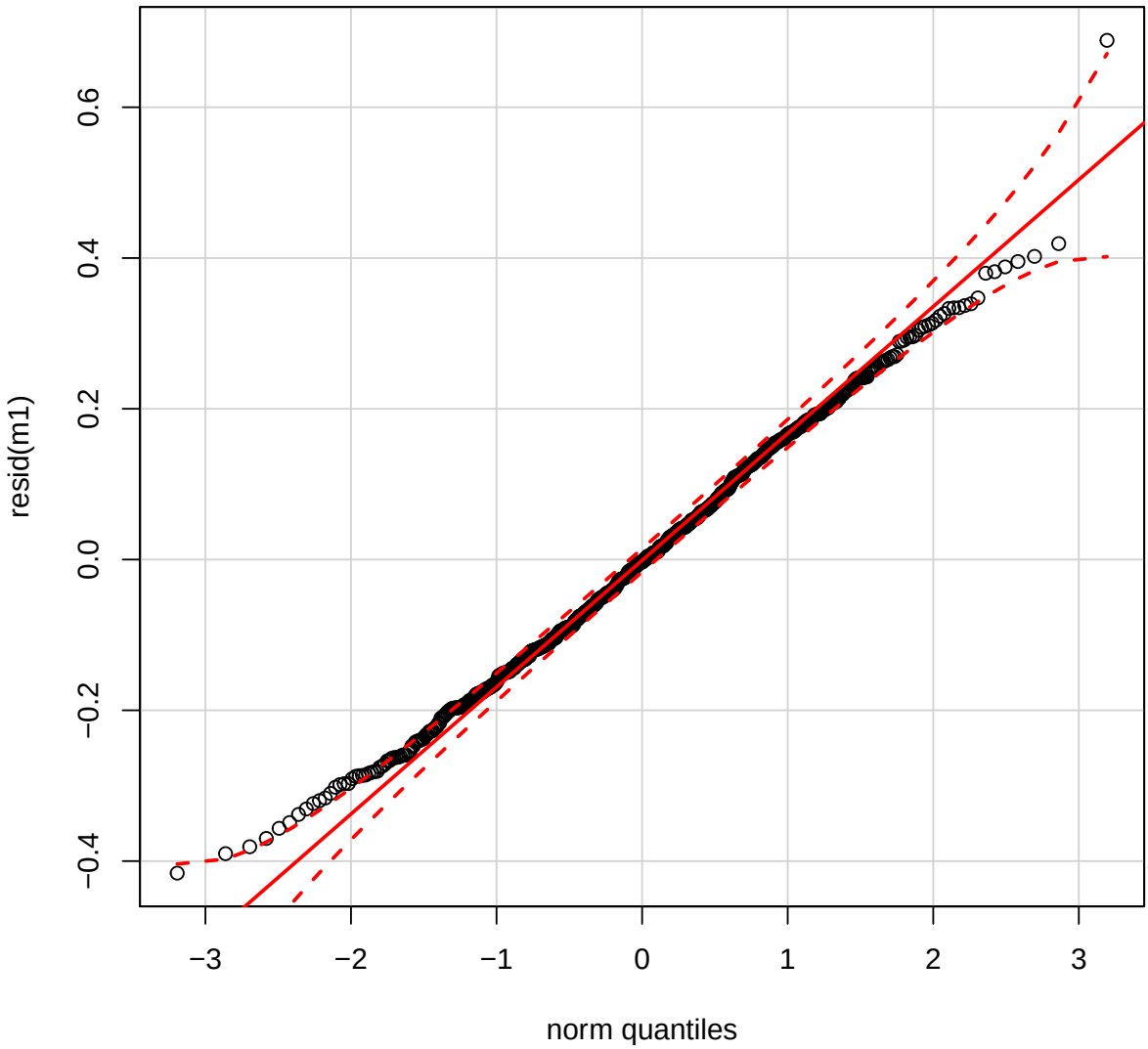
```
# gender distribution of participants
table(accents$IsMale)

##
##   0   1
## 329 383

round( 100*table(accents$IsMale)/nrow(accents), 1 )

##
##    0    1
## 46.2 53.8

# number of unique non-English languages in the dataset
length(unique(accents$Language))

## [1] 159

# average age of participants (and standard deviation)
mean(accents$Age)

## [1] 32.45716

sd(accents$Age)

## [1] 12.10157

# average age of English onset (and standard deviation)
mean(accents$AEO)

## [1] 12.17346

sd(accents$AEO)

## [1] 6.552217

# average length of residence in an English-speaking country (and standard deviation)
mean(accents$LR)

## [1] 6.823441

sd(accents$LR)

## [1] 10.74443

# correlation between the average number of education years per country and the
# gross national income
cor(accents$CountryEduYrs, accents$GNI.log)
```

```
## [1] 0.8363424

# correlation between Nativelikeness ratings and Levenshtein distances (LD)
# (note the correlation might be slightly different from the one reported
# by Wieling, Bloem, Mignella et al. (2014) as this study uses only
# non-native speakers who learned English academically)
cor(accents$Nativelikeness, accents$LD.log, use='pairwise')

## [1] -0.7636405

# correlation between LD-based speaker ratings (LD.log) and country distances (LDND)
cor(accents$LD.log,accents$LDND,use='pairwise')

## [1] 0.2994741

# correlation between speaker ratings (Nativelikeness) and country distances (LDND)
cor(accents$Nativelikeness,accents$LDND,use='pairwise')

## [1] -0.3704642

# number of speakers with Nativelikeness ratings
sum(!is.na(accents$Nativelikeness))

## [1] 237
```

3.2 Optimal model

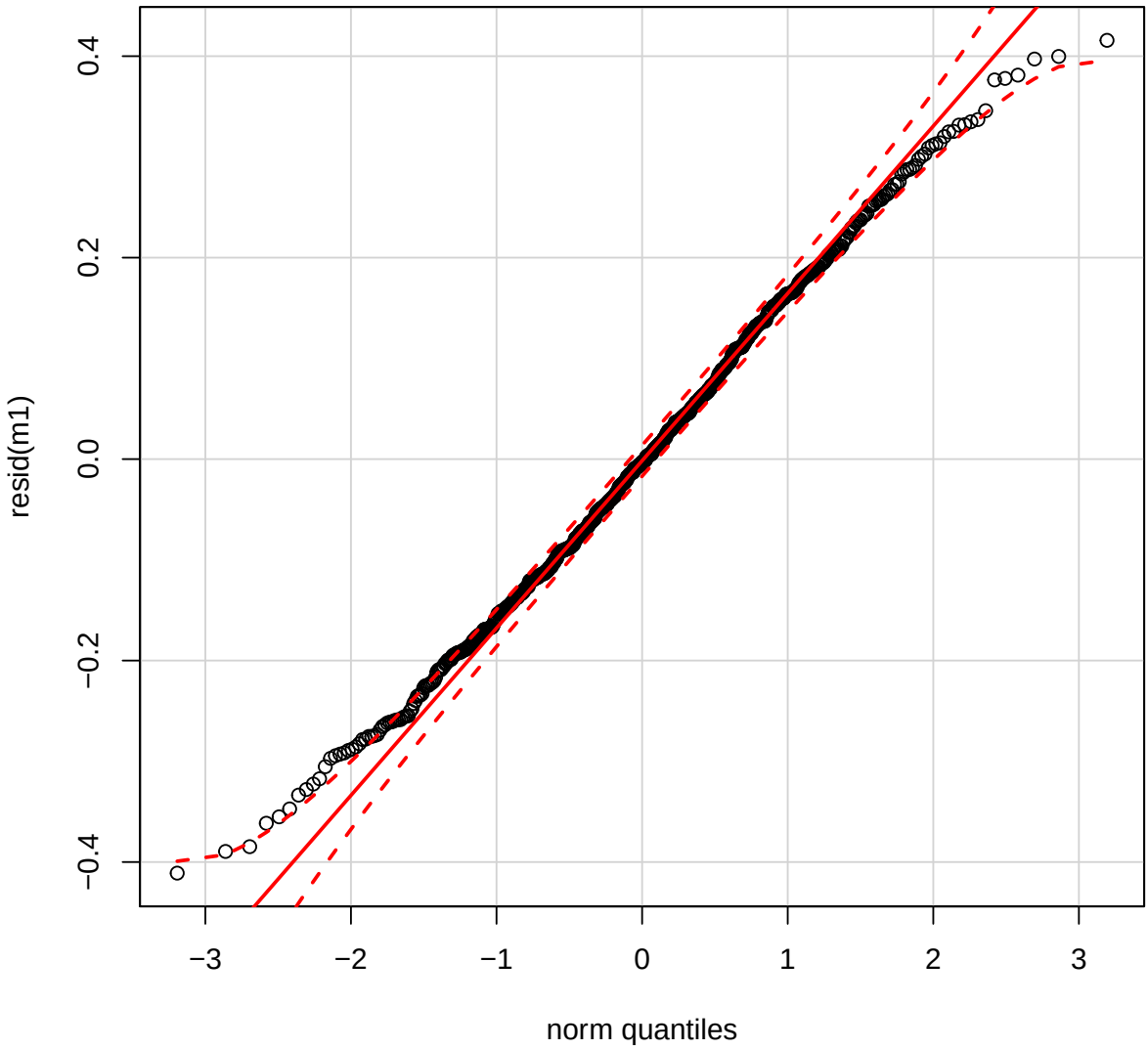
```
# optimal model specification determined via model comparison (not shown)
m0 <- gam(LD.log ~ LDND.c + AEO.c  + LR.c*Age.c+ CountryEduYrs.c + NrLang.c +
                   s(Country,bs='re') + s(Language,bs='re'), data=accents,
          method='REML')

summary(m0) # random intercept for Language not needed

##
## Family: gaussian
## Link function: identity
##
## Formula:
## LD.log ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c + NrLang.c +
##     s(Country, bs = "re") + s(Language, bs = "re")
##
## Parametric coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -4.815e+00  8.594e-03 -560.259  < 2e-16
## LDND.c           3.653e-03  9.102e-04    4.014 6.65e-05
## AEO.c            9.987e-03  1.033e-03    9.668  < 2e-16
## LR.c            -5.521e-03  1.046e-03   -5.279 1.76e-07
## Age.c            2.260e-03  6.899e-04    3.276 0.001108
```

```
## CountryEduYrs.c -1.964e-02  2.979e-03   -6.593 8.70e-11
## NrLang.c        -1.997e-02  5.833e-03   -3.424 0.000656
## LR.c:Age.c       1.046e-04  4.081e-05    2.564 0.010566
##
## Approximate significance of smooth terms:
##                   edf Ref.df     F  p-value
## s(Country)  3.193e+01    130 0.408 0.000359
## s(Language) 9.661e-04    157 0.000 0.561033
##
## R-sq.(adj) =  0.355   Deviance explained = 39.1%
## -REML = -221.6  Scale est. = 0.02621   n = 712

# final model is fit with ML, as in the GAM framework it offers
# more conservative p-values
m1 <- gam(LD.log ~ LDND.c + AEO.c  + LR.c*Age.c+ CountryEduYrs.c + NrLang.c +
                  s(Country,bs='re'), data=accents, method='ML')
summary(m1)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## LD.log ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c + NrLang.c +
##     s(Country, bs = "re")
##
## Parametric coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -4.815e+00  8.472e-03 -568.322  < 2e-16
## LDND.c           3.650e-03  9.011e-04    4.051 5.70e-05
## AEO.c            9.991e-03  1.032e-03    9.684  < 2e-16
## LR.c            -5.496e-03  1.044e-03   -5.265 1.89e-07
## Age.c            2.266e-03  6.894e-04    3.287 0.001065
## CountryEduYrs.c -1.958e-02  2.936e-03   -6.668 5.41e-11
## NrLang.c        -2.009e-02  5.824e-03   -3.450 0.000595
## LR.c:Age.c       1.037e-04  4.078e-05    2.542 0.011230
##
## Approximate significance of smooth terms:
##              edf Ref.df     F  p-value
## s(Country) 29.87    130 0.382 0.000388
##
## R-sq.(adj) =  0.353   Deviance explained = 38.7%
## -ML = -268.92  Scale est. = 0.026295  n = 712

qqp(resid(m1))
```




```
# 1 outlier at the top end is removed
accents2 = accents[resid(m1) < max(resid(m1)),]

# refitting without outlier
m1 <- gam(LD.log ~ LDND.c + AEO.c  + LR.c*Age.c+ CountryEduYrs.c + NrLang.c +
                   s(Country,bs='re'), data=accents2, method='ML')

summary(m1) # summary of final model

##
## Family: gaussian
```

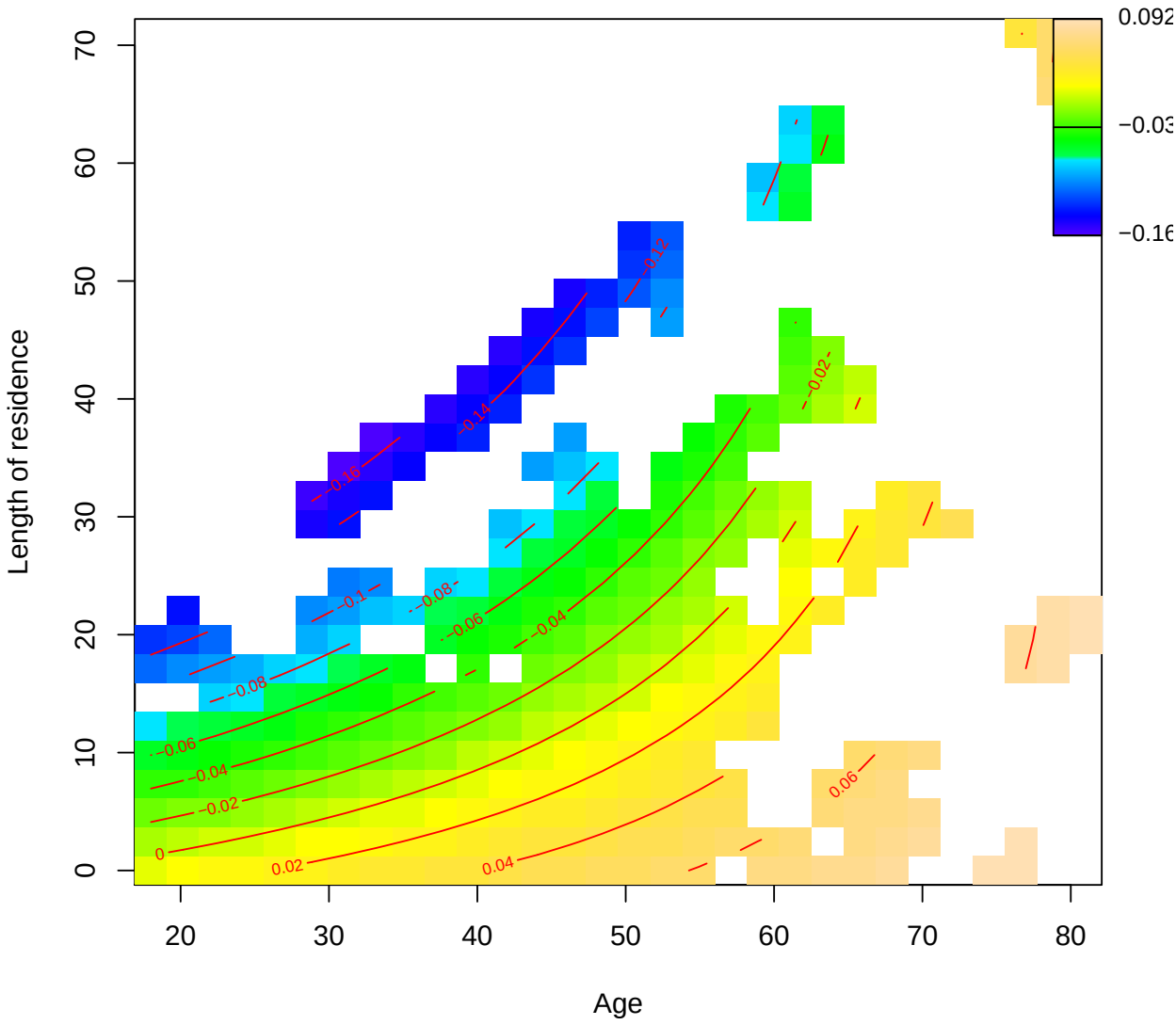
```
## Link function: identity
##
## Formula:
## LD.log ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c + NrLang.c +
##     s(Country, bs = "re")
##
## Parametric coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -4.816e+00  8.693e-03 -553.940  < 2e-16
## LDND.c           3.653e-03  9.134e-04    4.000 7.06e-05
## AEO.c            9.926e-03  1.020e-03    9.732  < 2e-16
## LR.c            -5.492e-03  1.034e-03   -5.313 1.47e-07
## Age.c            2.143e-03  6.808e-04    3.147 0.001723
## CountryEduYrs.c -1.971e-02  3.013e-03   -6.539 1.23e-10
## NrLang.c        -1.936e-02  5.763e-03   -3.359 0.000826
## LR.c:Age.c       1.077e-04  4.025e-05    2.675 0.007659
##
## Approximate significance of smooth terms:
##              edf Ref.df     F  p-value
## s(Country) 35.67    130 0.505 1.55e-05
##
## R-sq.(adj) =  0.368   Deviance explained = 40.6%
## -ML = -277.77  Scale est. = 0.025279  n = 711

qqp(resid(m1)) # residuals approximately normally distributed
```




```
# effect sizes:
coef(m1)["LDND.c"]*(max(accents2$LDND.c)-min(accents2$LDND.c))

##    LDND.c
## 0.2073974

coef(m1)["AEO.c"]*(max(accents2$AEO.c)-min(accents2$AEO.c))

##     AEO.c
## 0.4466788

coef(m1)["CountryEduYrs.c"]*(max(accents2$CountryEduYrs.c)-min(accents2$CountryEduYrs.c))
```

```
## CountryEduYrs.c
##      -0.2242543

coef(m1)["NrLang.c"]*(max(accents2$NrLang.c)-min(accents2$NrLang.c))

##    NrLang.c
## -0.09679126

mx = -Inf
mn = Inf
accents2$LR.c_Age.c = accents2$LR.c * accents2$Age.c
fx = coef(m1)
for (i in 1:nrow(accents2)) {
   val = fx["LR.c"]*accents2[i,"LR.c"] +
         fx["Age.c"]*accents2[i,"Age.c"] +
         fx["LR.c:Age.c"]*accents2[i,"LR.c_Age.c"]
   names(val) = "LR.c*Age.c"
   if (val < mn) {
      mn = val
   }
   if (val > mx) {
      mx = val
   }
}
round(mx - mn,3) # 0.24

## LR.c*Age.c
##       0.24

# explained variance without random intercept for country
summary(tmp <- gam(LD.log ~ LDND.c + AEO.c  + LR.c*Age.c+ CountryEduYrs.c + NrLang.c,
        data=accents2, method='ML'))

##
## Family: gaussian
## Link function: identity
##
## Formula:
## LD.log ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c + NrLang.c
##
## Parametric coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -4.813e+00  6.964e-03 -691.144  < 2e-16
## LDND.c           3.736e-03  7.828e-04    4.772 2.22e-06
## AEO.c            1.008e-02  1.012e-03    9.962  < 2e-16
## LR.c            -4.992e-03  1.018e-03   -4.906 1.15e-06
## Age.c            2.295e-03  6.843e-04    3.355 0.000838
## CountryEduYrs.c -1.809e-02  2.404e-03   -7.525 1.62e-13
## NrLang.c        -2.218e-02  5.690e-03   -3.898 0.000106
## LR.c:Age.c       8.907e-05  4.057e-05    2.195 0.028462
##
```

```
##
## R-sq.(adj) =  0.308   Deviance explained = 31.4%
## -ML = -270.39  Scale est. = 0.027677  n = 711

# visualization of interaction (refitting with tensor,
# which in this case simply models a linear interaction
# as the number of DF equals 3)
m1b <- gam(LD.log ~ LDND + AEO  + te(Age,LR)+ CountryEduYrs + NrLang +
                    s(Country,bs='re'), data=accents2, method='ML')

pvisgam(m1b, select=1, view=c('Age','LR'), too.far=0.04, hide.label=T,
        ylab='Length of residence', xlab='Age', main='')

## [1] "Tensor(s) to be plotted: te(Age,LR)"
```

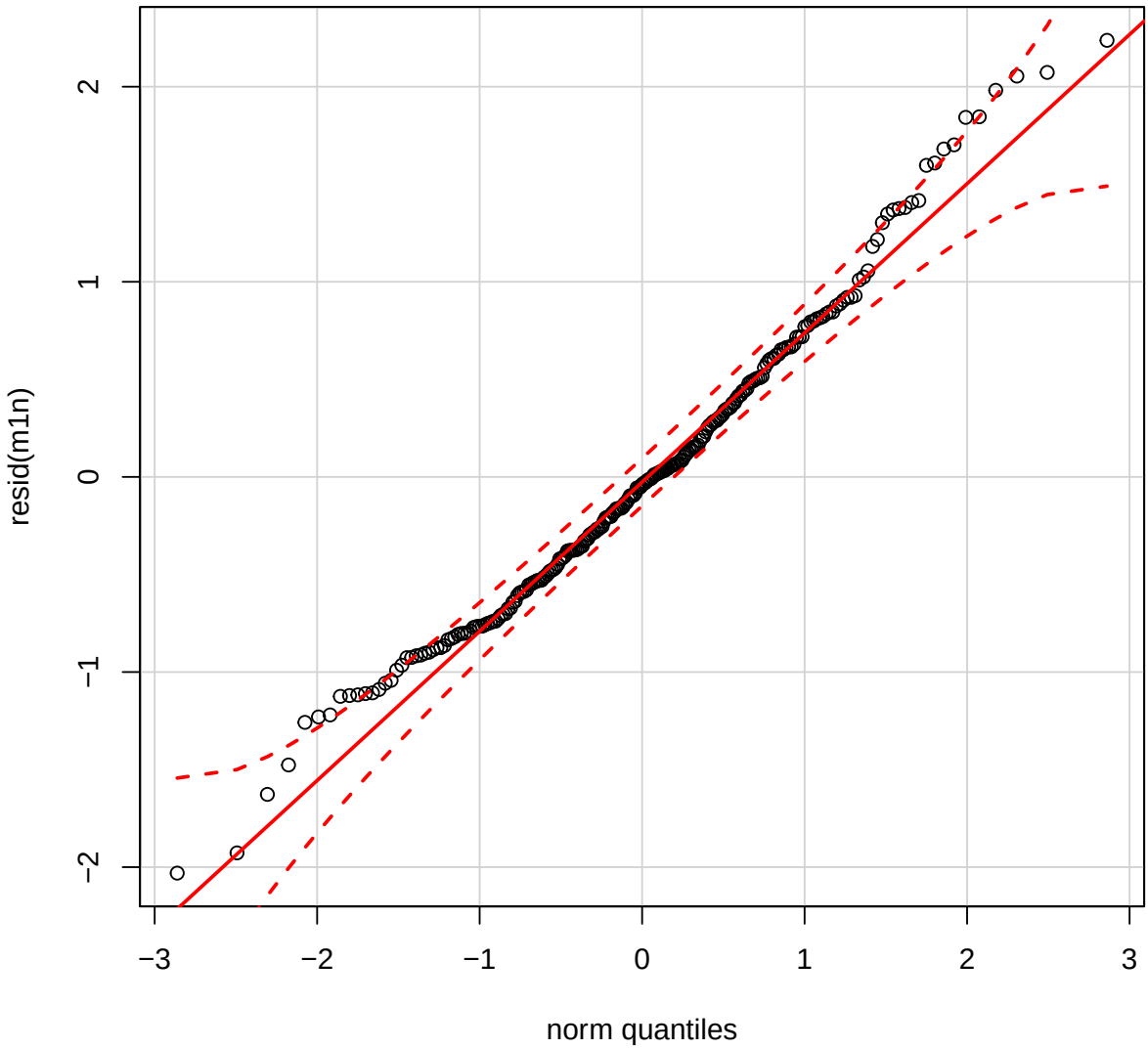


3.3 Valition with nativelikeness ratings

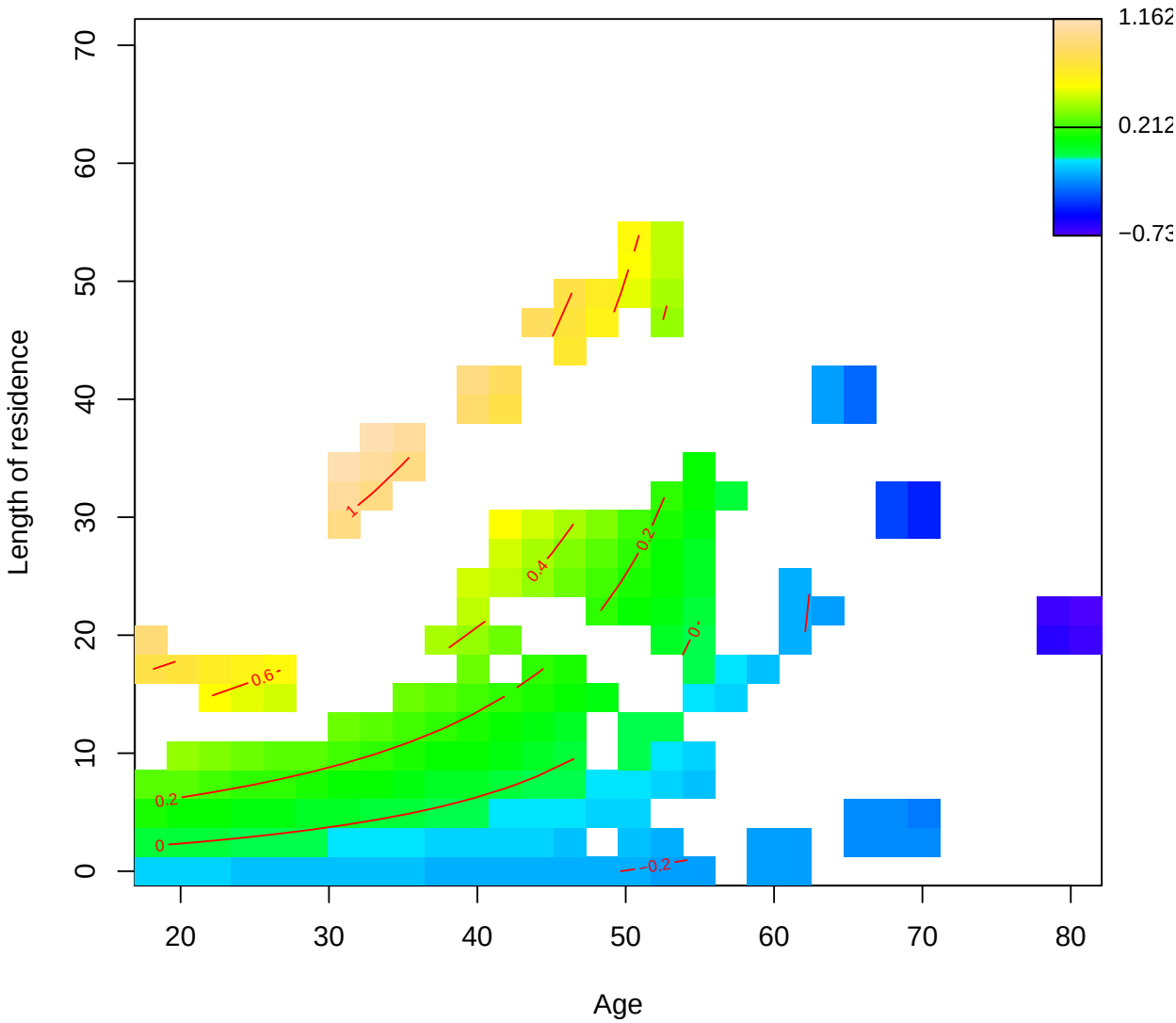
```
m0n <- gam(Nativelikeness ~ LDND.c + AEO.c  + LR.c*Age.c+ CountryEduYrs.c + NrLang.c +
                            s(Country,bs='re'), data=accents, method='ML')

summary(m0n) # random intercept for country not needed

##
## Family: gaussian
## Link function: identity
```

```
##
## Formula:
## Nativelikeness ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c +
##     NrLang.c + s(Country, bs = "re")
##
## Parametric coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.8268013  0.0642810  43.976  < 2e-16
## LDND.c          -0.0212581  0.0057234  -3.714 0.000256
## AEO.c           -0.0452705  0.0096076  -4.712 4.26e-06
## LR.c             0.0365949  0.0089080   4.108 5.56e-05
## Age.c           -0.0107250  0.0056574  -1.896 0.059255
## CountryEduYrs.c  0.0835559  0.0189305   4.414 1.57e-05
## NrLang.c         0.2839869  0.0456051   6.227 2.25e-09
## LR.c:Age.c      -0.0011767  0.0004874  -2.414 0.016551
##
## Approximate significance of smooth terms:
##               edf Ref.df     F p-value
## s(Country) 0.2968     80 0.004   0.364
##
## R-sq.(adj) =  0.392   Deviance explained = 41.1%
## -ML = 271.63  Scale est. = 0.59885   n = 237

m1n <- gam(Nativelikeness ~ LDND.c + AEO.c  + LR.c*Age.c+ CountryEduYrs.c + NrLang.c,
           data=accents, method='ML')

# summary of final model
summary(m1n)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## Nativelikeness ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c +
##     NrLang.c
##
## Parametric coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.8268615  0.0641851  44.042  < 2e-16
## LDND.c          -0.0212495  0.0057168  -3.717 0.000253
## AEO.c           -0.0452829  0.0096057  -4.714 4.22e-06
## LR.c             0.0365642  0.0089045   4.106 5.60e-05
## Age.c           -0.0107480  0.0056572  -1.900 0.058706
## CountryEduYrs.c  0.0834371  0.0188874   4.418 1.54e-05
## NrLang.c         0.2841480  0.0455951   6.232 2.19e-09
## LR.c:Age.c      -0.0011755  0.0004874  -2.412 0.016665
##
##
## R-sq.(adj) =  0.391   Deviance explained = 40.9%
## -ML = 271.63  Scale est. = 0.59972   n = 237
```

```r
# which is similar to the summary of the model
# on the basis of LD.log (but note inverted sign)
summary(m1)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## LD.log ~ LDND.c + AEO.c + LR.c * Age.c + CountryEduYrs.c + NrLang.c +
##     s(Country, bs = "re")
##
## Parametric coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     -4.816e+00  8.693e-03 -553.940  < 2e-16
## LDND.c           3.653e-03  9.134e-04    4.000 7.06e-05
## AEO.c            9.926e-03  1.020e-03    9.732  < 2e-16
## LR.c            -5.492e-03  1.034e-03   -5.313 1.47e-07
## Age.c            2.143e-03  6.808e-04    3.147 0.001723
## CountryEduYrs.c -1.971e-02  3.013e-03   -6.539 1.23e-10
## NrLang.c        -1.936e-02  5.763e-03   -3.359 0.000826
## LR.c:Age.c       1.077e-04  4.025e-05    2.675 0.007659
##
## Approximate significance of smooth terms:
##              edf Ref.df     F  p-value
## s(Country) 35.67    130 0.505 1.55e-05
##
## R-sq.(adj) =  0.368   Deviance explained = 40.6%
## -ML = -277.77  Scale est. = 0.025279  n = 711

# residuals are approximately normally distributed
qqp(resid(m1n))
```




```
# visualization of interaction (refitting with tensor,
# which in this case simply models a linear interaction
# as the number of DF equals 3)
m1nb <- gam(Nativelikeness ~ LDND + AEO  + te(Age,LR)+ CountryEduYrs + NrLang +
                             s(Country,bs='re'), data=accents2, method='ML')

pvisgam(m1nb, select=1, view=c('Age','LR'), too.far=0.04, hide.label=T,
        ylab='Length of residence', xlab='Age', main='')

## [1] "Tensor(s) to be plotted: te(Age,LR)"
```


Length of residence
Age
0
10
20
30
40
50
60
70
20
30
40
50
60
70
80
1.162
0.212
−0.73
1
0.6
0.4
0.2
0
0.2
−0.2