Case study: null subjects

Gemma Hunter McCarley

Contents

1 Load dependencies and data

```
library(tidyverse)
library(lme4)
library(DHARMa)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

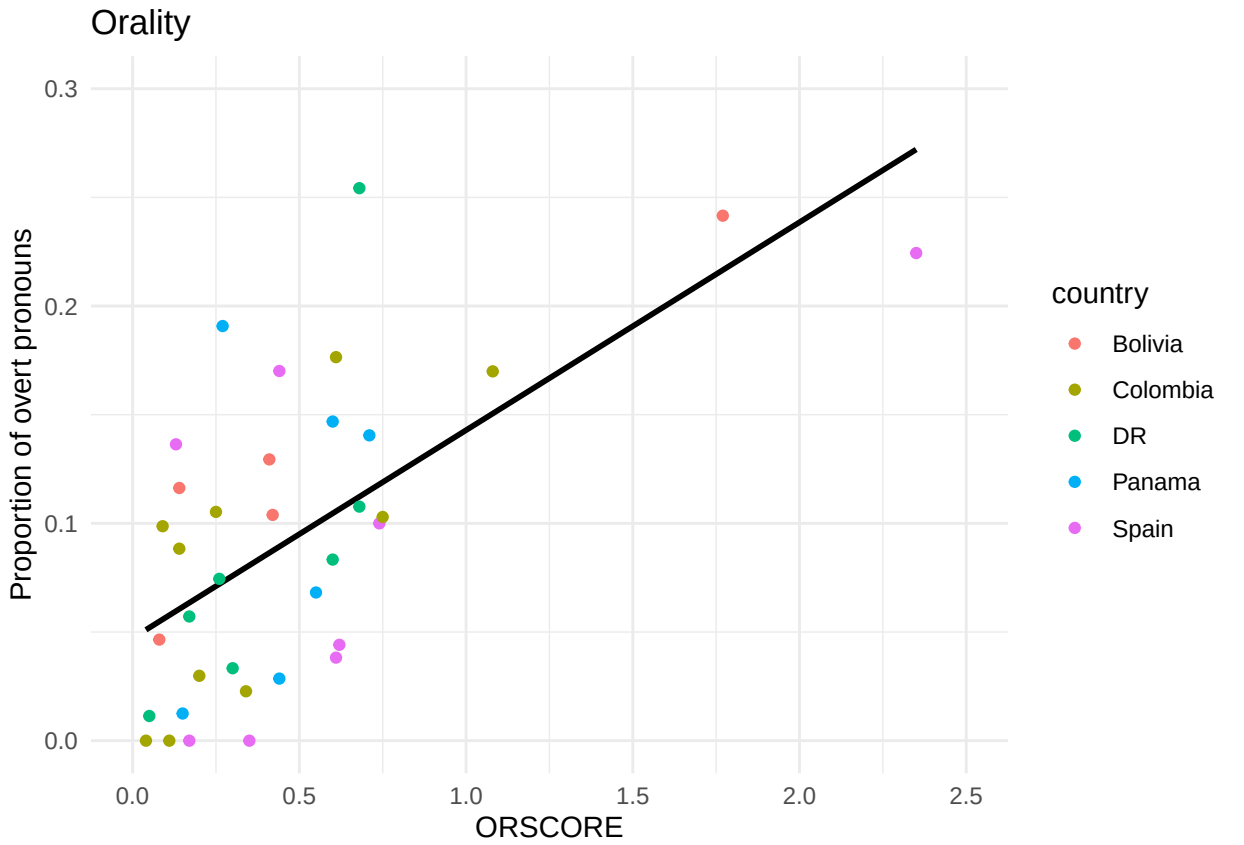
```
## The following object is masked from 'package:purrr':
##
##     some
```

```
mydata <- read.csv('cordeles_2023.csv', header = TRUE, encoding = "UTF-8")
orality <- read.csv('orality.csv', header = TRUE, encoding = "UTF-8")

#Orality####
orality <- subset(orality, docID != "CPMTpoSP") #get rid of duplicate data (translation of another text
```

2 Plots and models

```
#plot
orality_plot_regression <- ggplot(data = orality, aes(x = ORSCORE, y = OVERT_RATE)) +
  ylim(0, .3) +
  xlim(0, 2.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(x="ORSCORE", y="Proportion of overt pronouns") +
  theme_minimal()
orality_plot_regression + geom_point(aes(color = country)) + ggtitle(paste0("Orality")) + theme_minimal
```

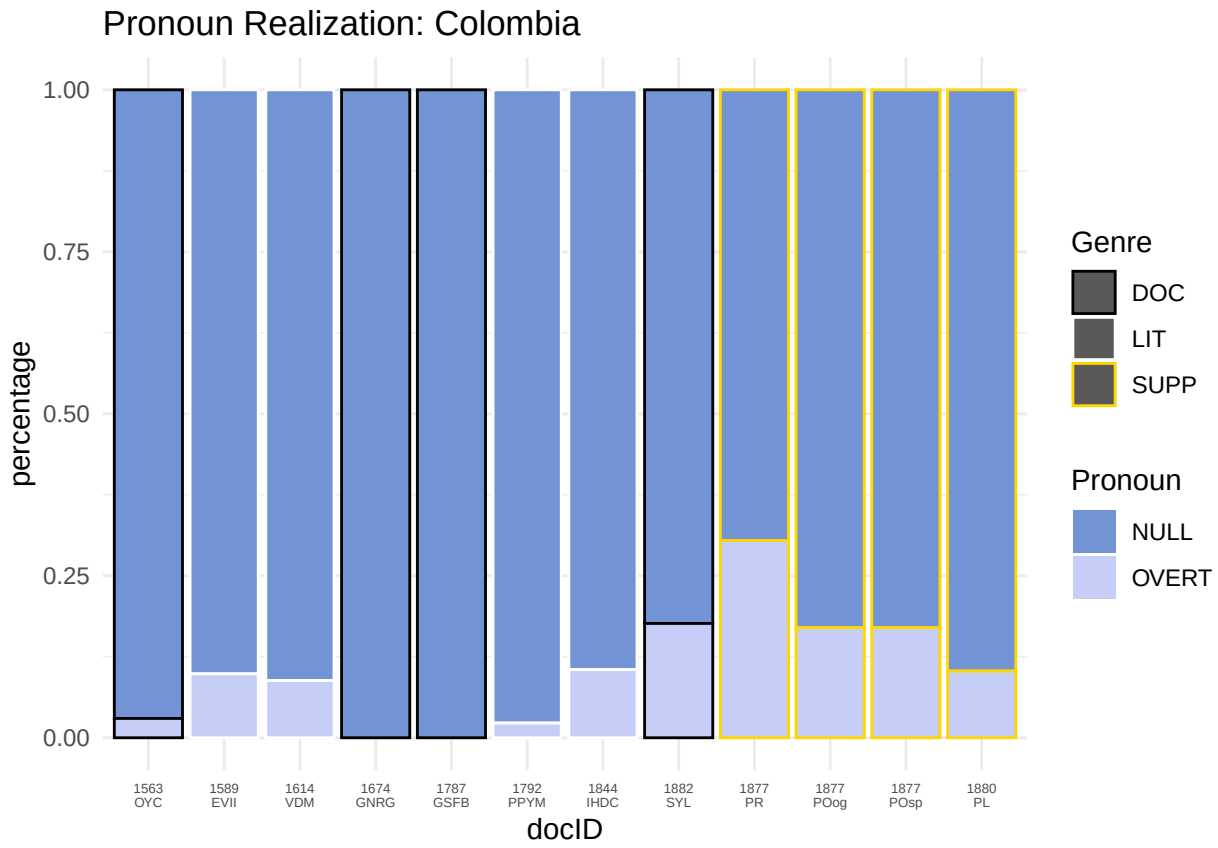
```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



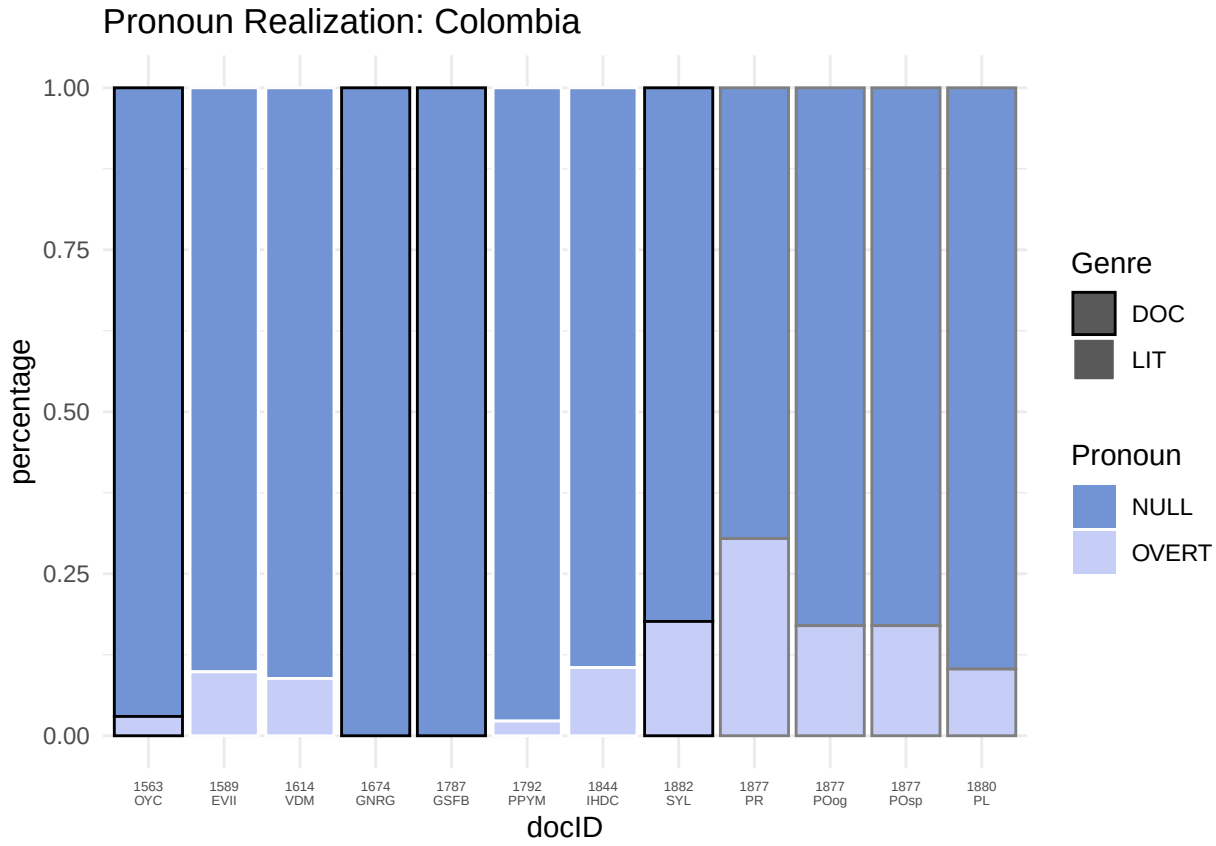

```
#model
xmdl <- lm(OVERT_RATE ~ ORSCORE, data = orality)
summary(xmdl)
```

```
##
## Call:
## lm(formula = OVERT_RATE ~ ORSCORE, data = orality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08527 -0.05271 -0.01067  0.03651  0.18698
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.05016    0.01537   3.263 0.002465 **
## ORSCORE      0.10030    0.02307   4.348 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06403 on 35 degrees of freedom
## Multiple R-squared:  0.3507, Adjusted R-squared:  0.3322
## F-statistic: 18.91 on 1 and 35 DF,  p-value: 0.0001128
```

```r
#Pronoun Plots####
#make country specific (country, re-leveling, and year labels):
country <- "Colombia" #change for new country
data <- filter(mydata, Country == country) # change to "DR" for Dominican Republic

#un-comment the one you need below
data$docID <- factor(data$docID,levels = c("co16oyc", "co16evii", "co17vdm", "co17gnrg", "co18gsfb", "c
#data$docID <- factor(data$docID,levels = c("dr16sdj", "dr16ent", "dr17dphj", "dr18asd", "dr18livie", "
#data$docID <- factor(data$docID,levels = c("bo16rvp", "bo18hvip", "bo19adla", "bo19jdlr", "bo21iia"))
#data$docID <- factor(data$docID,levels = c("pa16hgni", "pa16car", "pa17lldp", "pa17dlyd", "pa19mpe", "
#data$docID <- factor(data$docID,levels = c("sp16can", "sp16lah", "sp17dq", "sp17acra", "sp18arjd", "sp

#un-comment the one you need below
docs <- c("1563\nOYC", "1589\nEVII", "1614\nVDM", "1674\nGNRG", "1787\nGSFB", "1792\nPPYM", "1844\nIHDC
#docs <- c("1500\nSDJ", "1588\nENT", "1658\nDPHJ", "1752\nASD", "1785\nLIVIE", "1857\nALD", "1886\nGAL"
#docs <- c("1550\nRVP", "1721\nHVIP", "1839\nADLA", "1885\nJDLR", "2010\nIIA")
#docs <- c("1535\nHGNI", "1546\nCAR", "1638\nLLDP", "1695\nDLYD", "1872\nMPE", "1875\nHS")
#docs <- c("1525\nCAN", "1551\nLAH", "1605\nDQ", "1664\nACRA", "1756\nARJD", "1786\nEAU", "1836\nQDEV",

#plot (repeat for each country)
barchart <- data %>%
  ggplot(aes(x = docID, fill = sub_POS)) +
  scale_x_discrete(labels=docs) +
  geom_bar(position = "fill", aes(color=Genre), linewidth = .5) +
  ylab("percentage") +
  guides(fill=guide_legend(title="Pronoun"), color=guide_legend(title="Genre")) +
  ggtitle(paste0("Pronoun Realization: ", country))+
  theme_minimal()

barchart + scale_fill_manual(values = c("NULL" = "#7294D4", "OVERT" = "#C6CDF7")) + scale_color_manual(
```




```
#supp genre ^^^

barchart + scale_fill_manual(values = c("NULL" = "#7294D4", "OVERT" = "#C6CDF7")) + scale_color_manual(
```




```
#regular ^^^

#Main Model####
#Get data ready
binary_null <- within(mydata, Country <- relevel(factor(Country), ref = "Spain")) #make Spain the refer
binary_null <- within(binary_null, Macro_Region <- relevel(factor(Macro_Region), ref = "Spain")) #make

binary_null %>% select(Region, Genre, sub_POS, Country, Year, ORSCORE, docID, Century, Macro_Region) %>%
  na.omit() %>%
  mutate(Century = as.factor(Century)) -> binary_null

binary_null <- subset(binary_null, binary_null$docID != "CPMTpoSP") #get rid of duplicate data (transla

#Model
binary_null_mdl_int <- glmer(factor(sub_POS) ~ scale(Year) * scale(ORSCORE) + Macro_Region + (1|docID),
                             data = binary_null,
                             family = 'binomial')
```

3 Tests

```
#tests
summary(binary_null_mdl_int)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
```

```
## Formula: factor(sub_POS) ~ scale(Year) * scale(ORSCORE) + Macro_Region +
##     (1 | docID)
##    Data: binary_null
##
##      AIC      BIC   logLik deviance df.resid
##   2548.2   2585.6  -1268.1   2536.2     3767
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.5568 -0.4031 -0.2986 -0.2208  7.2692
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  docID  (Intercept) 0.3119   0.5585
## Number of obs: 3773, groups:  docID, 37
##
## Fixed effects:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.54816    0.27570  -9.242  < 2e-16 ***
## scale(Year)                 0.05029    0.15364   0.327 0.743434
## scale(ORSCORE)              1.02970    0.26917   3.825 0.000131 ***
## Macro_RegionNon-Spain       0.62880    0.31997   1.965 0.049389 *
## scale(Year):scale(ORSCORE) -0.43702    0.20634  -2.118 0.034177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) scl(Y) s(ORSC M_RN-S
## scale(Year)  0.086
## sc(ORSCORE) -0.053 -0.577
## Mcr_RgnNn-S -0.831 -0.307  0.388
## s(Y):(ORSCO -0.180  0.356 -0.723 -0.170
```

```
drop1(binary_null_mdl_int, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## factor(sub_POS) ~ scale(Year) * scale(ORSCORE) + Macro_Region +
##     (1 | docID)
##                            npar    AIC    LRT Pr(Chi)
## <none>                          2548.2
## Macro_Region                  1 2550.0 3.8159 0.05077 .
## scale(Year):scale(ORSCORE)    1 2550.6 4.4147 0.03563 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(binary_null_mdl_int, type = "III")
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: factor(sub_POS)
##                             Chisq Df Pr(>Chisq)
## (Intercept)                85.4228  1  < 2.2e-16 ***
## scale(Year)                 0.1071  1  0.7434341
```

```
## scale(ORSCORE)             14.6342  1  0.0001305 ***
## Macro_Region                3.8621  1  0.0493893 *
## scale(Year):scale(ORSCORE)  4.4858  1  0.0341767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(binary_null_mdl_int)
```

```
## [1] 2548.202
```

```
#checking residuals
res <- simulateResiduals(binary_null_mdl_int, plot = T)
```