Codes for Final Project (ver 3.0)

Xinyao Yi

Installing Packages and Library

```
library(tidyverse)
library(lme4)
library(data.table)
library(stringr)
library(dplyr)
library(graphics)
library(lmerTest)
library(ggeffects)
```

Load Dataset

```
load("final_data.RData")
```

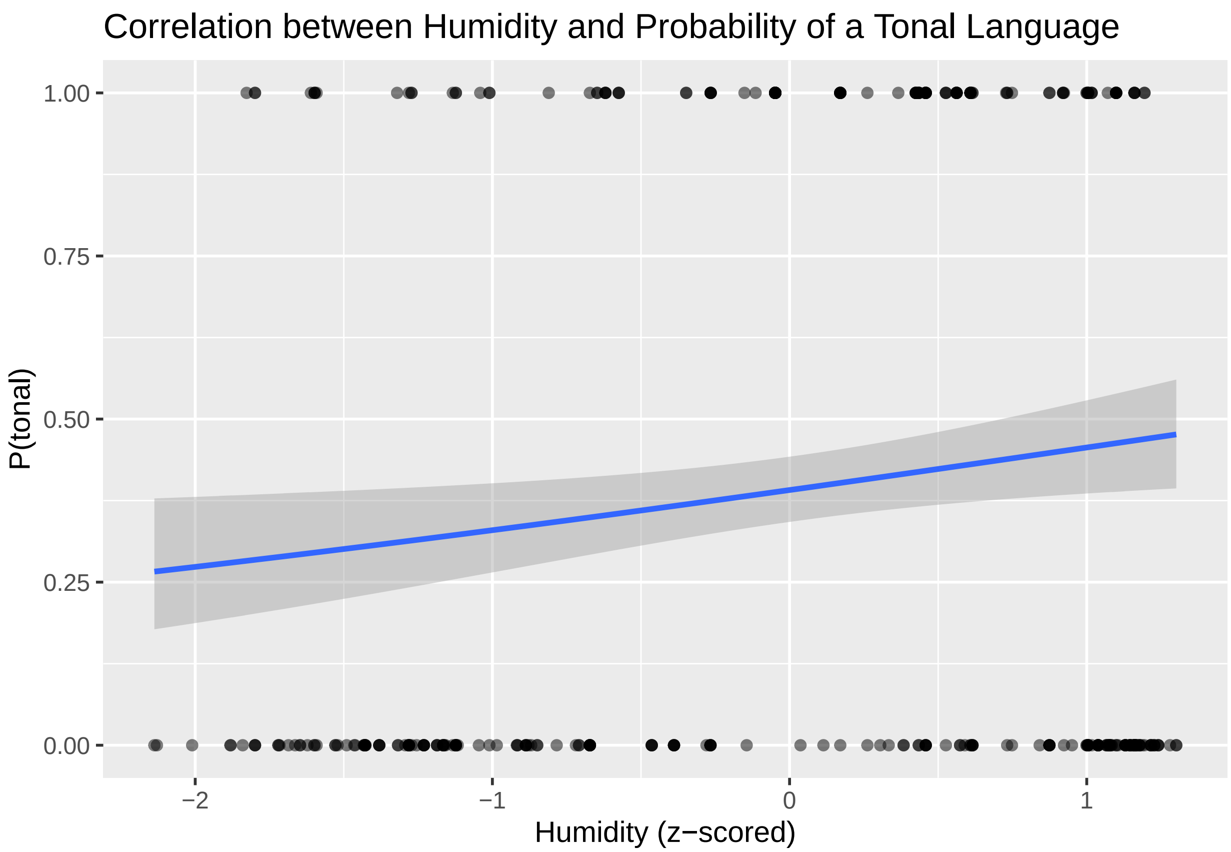
Data preprocessing

```
#Part 1: Change `tone system` into a binomial variable
##tone_num: dummy coding string variable `tones` to a 3-level numeric variable `tone_num`
final.data$tone_num = final.data$tones

final.data$tone_num = ifelse(final.data$tone_num  == '1 - No tones', 1,
                              ifelse(final.data$tone_num  == '2 - Simple tone system', 2,
                                 ifelse(final.data$tone_num  == '3 - Complex tone system', 3,
                                                          1)))
final.data = final.data %>%
  mutate(tone_num = as.numeric(tone_num))

##tone_bin: make 3-level variable `tone_num` into a binomial variable `tone_bin`
final.data$tone_bin = final.data$tone_num

final.data$tone_bin = ifelse(final.data$tone_bin  == 1, 0,
                              ifelse(final.data$tone_bin == 2, 1,
                                 ifelse(final.data$tone_bin == 3, 1,
                                                          0)))

final.data$tone_bin = as.numeric(final.data$tone_bin)
```

```r
#Part 2: Take z-score of environmental features (humidity and temperature)
#formula: z_scores <- (data - mean(data)) / sd(data)
mean_hum_repl = mean(final.data$mean_hum)
sd_hum_repl = sd(final.data$mean_hum)

mean_elev_repl = mean(final.data$elev_m)
sd_elev_repl = sd(final.data$elev_m)

final.data = final.data %>%
  mutate(humidity_z = (mean_hum - mean_hum_repl) / sd_hum_repl) %>%
  mutate(elevation_z = (elev_m - mean_elev_repl) / sd_elev_repl)

#Part 3: Handling outliers (3 outliers are dropped from the data set)
final.data = final.data %>%
  filter(between(humidity_z, -5, 5)) %>%
  filter(between(elevation_z, -5, 5))
```

Data Visualization

```r
ggplot(data = final.data,
       mapping = aes(x = humidity_z,
                     y = tone_bin)) +
  geom_point(alpha=.5) +
  geom_smooth(method="glm",
              method.args = list(family = "binomial")) +
  labs(x = "Humidity (z-scored)",
       y = "P(tonal)") +
  ggtitle("Correlation between Humidity and Probability of a Tonal Language")
```



Do models

```
#Compact model: tone ~ humidity
tone_hum = glm(data = final.data, tone_bin ~ 1 + humidity_z, family = "binomial")
summary(tone_hum)
```

```
## 
## Call:
## glm(formula = tone_bin ~ 1 + humidity_z, family = "binomial", 
##     data = final.data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.1377  -1.0602  -0.8634   1.3037   1.5896  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -0.4423     0.1074  -4.118 3.82e-05 ***
## humidity_z    0.2675     0.1094   2.446   0.0144 *  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
```

```
##     Null deviance: 497.36  on 370  degrees of freedom
## Residual deviance: 491.24  on 369  degrees of freedom
## AIC: 495.24
##
## Number of Fisher Scoring iterations: 4
```

```
#Augmented model: tone ~ humidity + elevation
tone_hum_elev = glm(data = final.data, tone_bin ~ 1 + humidity_z + elevation_z, family = "binomial")
summary(tone_hum_elev)
```

```
##
## Call:
## glm(formula = tone_bin ~ 1 + humidity_z + elevation_z, family = "binomial",
##     data = final.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4386  -1.0485  -0.8274   1.3039   1.6642
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4361     0.1079  -4.042 5.29e-05 ***
## humidity_z    0.2963     0.1118   2.651  0.00802 **
## elevation_z   0.2118     0.1258   1.683  0.09238 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 497.36  on 370  degrees of freedom
## Residual deviance: 488.40  on 368  degrees of freedom
## AIC: 494.4
##
## Number of Fisher Scoring iterations: 4
```

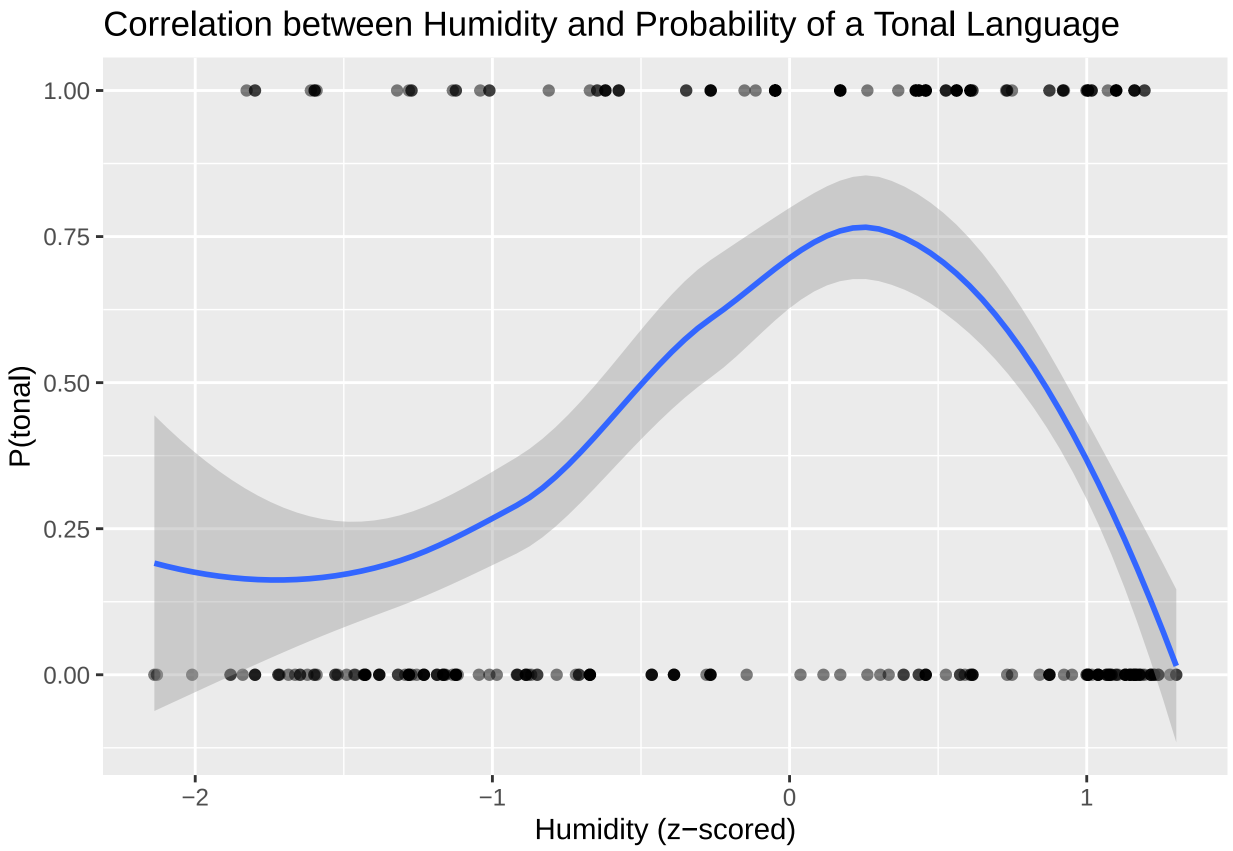
Interpret results: Calculate probability

```
#Compact model: tone ~ humidity
#(value range of humidity_z: [-2.13, 1.30])
ggpredict(model = tone_hum,
          terms = "humidity_z [-3:2]")
```

```
## # Predicted probabilities of tone_bin
##
## humidity_z | Predicted |       95% CI
## --------------------------------------
##         -3 |      0.22 | [0.13, 0.36]
##         -2 |      0.27 | [0.19, 0.38]
##         -1 |      0.33 | [0.26, 0.40]
##          0 |      0.39 | [0.34, 0.44]
##          1 |      0.46 | [0.39, 0.53]
##          2 |      0.52 | [0.41, 0.64]
```

Discussion 2: Graph

```
ggplot(data = final.data,
       mapping = aes(x = humidity_z,
                     y = tone_bin)) +
  geom_point(alpha=.5) +
  geom_smooth() +
  labs(x = "Humidity (z-scored)",
       y = "P(tonal)") +
  ggtitle("Correlation between Humidity and Probability of a Tonal Language")
```



```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
```

```
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggeffects_1.1.4   lmerTest_3.1-3    data.table_1.14.2 lme4_1.1-30
##  [5] Matrix_1.5-1      forcats_0.5.2     stringr_1.4.1     dplyr_1.0.10
##  [9] purrr_0.3.4       readr_2.1.2       tidyr_1.2.1       tibble_3.1.8
## [13] ggplot2_3.4.0     tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.4          jsonlite_1.8.0      splines_4.2.1
##  [4] modelr_0.1.9        assertthat_0.2.1    highr_0.9
##  [7] googlesheets4_1.0.1 cellranger_1.1.0    yaml_2.3.5
## [10] numDeriv_2016.8-1.1 pillar_1.8.1        backports_1.4.1
## [13] lattice_0.20-45     glue_1.6.2          digest_0.6.29
## [16] snakecase_0.11.0    rvest_1.0.3         minqa_1.2.5
## [19] colorspace_2.0-3    htmltools_0.5.3     pkgconfig_2.0.3
## [22] broom_1.0.1         haven_2.5.1         scales_1.2.1
## [25] tzdb_0.3.0          googledrive_2.0.0   mgcv_1.8-40
## [28] generics_0.1.3      farver_2.1.1        sjlabelled_1.2.0
## [31] ellipsis_0.3.2      withr_2.5.0         cli_3.4.1
## [34] magrittr_2.0.3      crayon_1.5.1        readxl_1.4.1
## [37] evaluate_0.16       fs_1.5.2            fansi_1.0.3
## [40] nlme_3.1-157        MASS_7.3-58.1       xml2_1.3.3
## [43] tools_4.2.1         hms_1.1.2           gargle_1.2.1
## [46] lifecycle_1.0.3     munsell_0.5.0       reprex_2.0.2
## [49] compiler_4.2.1      rlang_1.0.6         grid_4.2.1
## [52] nloptr_2.0.3        rstudioapi_0.14     labeling_0.4.2
## [55] rmarkdown_2.16      boot_1.3-28         gtable_0.3.1
## [58] DBI_1.1.3           R6_2.5.1            lubridate_1.8.0
## [61] knitr_1.40          fastmap_1.1.0       utf8_1.2.2
## [64] insight_0.18.6      stringi_1.7.8       Rcpp_1.0.9
## [67] vctrs_0.5.1         dbplyr_2.2.1        tidyselect_1.1.2
## [70] xfun_0.33
```