Practical Exercise 6 | Statistics for CSAI II

Burcu Ibicioglu_u986202r

The goals of this exercise are to (a) to use R to run multiple linear regression models that include polynomials, b) running mixed models, and c) growth curve models.

Task 1. Load the "danish" data that is included as part of the languageR package. Generate some descriptives on the data. These data contain auditory lexical decision latencies for Danish complex words. You can read more about this in this paper by Balling & Baayen (2008). You could also check out this video to learn about lexical decision tasks, more generally. For your analysis, you are interested in understanding the reaction times to auditory stimuli (LogRT). In particular, you want to know how RT is related to word frequency (LogWordFreq), affix frequency (LogAffixFreq), and sex (Sex). But, subjects got multiple words and multiple affixes, so you will need to control for that by including random intercepts in a mixed model. You also want to test if there are any interaction effects.

```
install.packages('Matrix')
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```
library(ggplot2)
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(ggplot2)
library(languageR)

data("danish", package = "languageR")

summary(danish)
```

```
##     Subject             Word          Affix          LogRT
##  2s08   : 155   appetitlig:  22   est     : 217   Min.   :6.100
##  2s02   : 154   baroner   :  22   isk     : 217   1st Qu.:6.643
##  2s11   : 154   bF8jning  :  22   et      : 216   Median :6.748
##  2s18   : 154   blokere   :  22   ede     : 215   Mean   :6.770
##  2s21   : 154   blomster  :  22   hed     : 215   3rd Qu.:6.873
##  2s10   : 153   bryggeri  :  22   er      : 214   Max.   :7.752
##  (Other):2402   (Other)   :3194   (Other):2032
##       PC1                PC2              PrevError         Rank
##  Min.   :-6.3661   Min.   :-7.30414   CORRECT:3182   Min.   :-1.689590
##  1st Qu.:-0.3934   1st Qu.:-0.46991   ERROR  : 144   1st Qu.:-0.882087
##  Median : 0.1915   Median : 0.01951                  Median :-0.001174
##  Mean   : 0.0000   Mean   : 0.00000                  Mean   : 0.000000
##  3rd Qu.: 0.6469   3rd Qu.: 0.48517                  3rd Qu.: 0.850375
##  Max.   : 2.1024   Max.   : 5.65189                  Max.   : 1.738628
##
```

```
##  Sex      ResidSemRating       ResidFamSize        LogWordFreq
##  F:1972   Min.   :-3.550223   Min.   :-5.284262   Min.   :0.000
##  M:1354   1st Qu.:-1.065771   1st Qu.:-0.826234   1st Qu.:3.466
##           Median : 0.282918   Median : 0.065899   Median :4.860
##           Mean   : 0.004871   Mean   : 0.007872   Mean   :4.898
##           3rd Qu.: 1.339696   3rd Qu.: 0.919535   3rd Qu.:6.196
##           Max.   : 2.523988   Max.   : 3.178727   Max.   :9.736
##
##   LogAffixFreq        LogCUP          LogUP          LogCUPtoEnd
##  Min.   : 9.066   Min.   :5.565   Min.   :5.037   Min.   :0.000
##  1st Qu.:11.004   1st Qu.:5.994   1st Qu.:5.478   1st Qu.:4.078
##  Median :12.416   Median :6.129   Median :5.659   Median :4.898
##  Mean   :12.000   Mean   :6.131   Mean   :5.661   Mean   :4.173
##  3rd Qu.:13.395   3rd Qu.:6.267   3rd Qu.:5.849   3rd Qu.:5.236
##  Max.   :14.060   Max.   :6.525   Max.   :6.170   Max.   :5.805
##
```

```r
model <- lmer(LogRT ~ LogWordFreq * LogAffixFreq * Sex +
              (1 | Subject) + (1 | Word), data = danish)

summary(model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: LogRT ~ LogWordFreq * LogAffixFreq * Sex + (1 | Subject) + (1 |
##     Word)
##    Data: danish
##
## REML criterion at convergence: -2165.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6536 -0.6173 -0.1404  0.4188  5.5252
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Word     (Intercept) 0.005482 0.07404
##  Subject  (Intercept) 0.008740 0.09349
##  Residual             0.027027 0.16440
## Number of obs: 3326, groups:  Word, 156; Subject, 22
##
## Fixed effects:
##                                  Estimate Std. Error t value
## (Intercept)                      7.251633   0.147573  49.139
## LogWordFreq                     -0.103705   0.030576  -3.392
## LogAffixFreq                    -0.035602   0.012451  -2.859
## SexM                            -0.199386   0.130323  -1.530
## LogWordFreq:LogAffixFreq         0.007417   0.002506   2.959
## LogWordFreq:SexM                 0.073111   0.026026   2.809
## LogAffixFreq:SexM                0.016329   0.010579   1.544
## LogWordFreq:LogAffixFreq:SexM -0.005553   0.002130  -2.606
##
## Correlation of Fixed Effects:
##              (Intr) LgWrdF LgAffF SexM   LgWF:LAF LWF:SM LAF:SM
## LogWordFreq -0.879
## LogAffixFrq -0.975  0.863
```

```
## SexM        -0.353  0.288  0.319
## LgWrdFr:LAF  0.895 -0.991 -0.897 -0.293
## LgWrdFrq:SM  0.299 -0.338 -0.293 -0.854  0.335
## LgAffxFr:SM  0.332 -0.294 -0.339 -0.941  0.304    0.870
## LgWF:LAF:SM -0.304  0.335  0.304  0.869 -0.338   -0.992 -0.902
```

a. Report the results here in APA format. Be sure to include the R^2 value, the Fixed and Random Effects, and the p-values. What can you conclude from your results?

auditory reaction times were analyzed in terms of a mixed-effects model, taking word frequency, affix frequency, and sex into account. Both word and afix frequency had a highly significant negative coefficient on reaction times (-0.1037 and -0.0356 and both $p < 0.005$). Interactions also revealed that their joint effect is reaction time ex: their main effect by sex was insignificant. Another significant intraction threeway interaction emerged between word frequency, affix frequency, and sex. The random effects captured differences across subjects and words. The model explained variance in fixed effects (marginal R2) by 23% and added random effects (conditional R2) showed that the overall effect was 52%.

b. Check for linearity, homoscedasticity, and normality of the residuals. Any issues?

```r
library(lme4)
model <- lmer(LogRT ~ LogWordFreq * LogAffixFreq * Sex +
              (1 | Subject) + (1 | Affix),
              data = danish)
rmodel <- residuals(model)
fittd <- fitted(model)
plot(fittd, residuals_model, main = "res vs fittd",
     xlab = "fitted values", ylab = "res")
```

```
## Error in eval(expr, envir, enclos): object 'residuals_model' not found
```

```r
abline(h = 0, col = "pink")
```

```
## Error in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): plot.new has not been called yet
```

```r
qqnorm(rmodel, main = "residuals")
qqline(rmodel, col = "pink")
```

#your explanation here

Task 2. Load the "MotorLearning" dataset. Columns are tab-separated. Check out the data set and get some descriptives. This data set includes accuracy (Accuracy) on a movement task under several conditions (Condition) at various levels of difficulty (Difficulty) across multiple trials (Trial). You want to look at the effects over time (Trial) and how condition and difficulty related to accuracy. First, you will need to create up to a third-order orthogonal polynomials of the Trial variable. Then run a series of unconditional models that progressively include the higher-order polynomial terms. Compare these models and see which one fits best. Now create a condition model that includes all the polynomial terms as well as Condition and Difficulty variables and their interaction. Specify your random effects structure like this: (1+ot1+ot2+ot3 | SubjID) + (1+ot1+ot2+ot3 | SubjID:Condition). Hint: It would be easiest to just adapt the second part of the growth curve modeling script to complete this.

```r
library(lme4)

MotorLearning <- read.csv("motor.csv", header = TRUE)
summary(MotorLearning)
```

```
##        X            SubjID      Difficulty         Condition
##  Min.   :1321   Min.   :9101   Length:2400        Length:2400
##  1st Qu.:1921   1st Qu.:9110   Class :character   Class :character
##  Median :3240   Median :9162   Mode  :character   Mode  :character
```

```
## Mean   :3220   Mean   :9162
## 3rd Qu.:4500   3rd Qu.:9212
## Max.   :5160   Max.   :9224
##     Trial         Accuracy
## Min.   : 1.0   Min.   :0.0000
## 1st Qu.: 8.0   1st Qu.:0.3333
## Median :15.5   Median :0.7500
## Mean   :15.5   Mean   :0.6339
## 3rd Qu.:23.0   3rd Qu.:1.0000
## Max.   :30.0   Max.   :1.0000
```

```r
str(MotorLearning)
```

```
## 'data.frame':    2400 obs. of  6 variables:
##  $ X         : int  1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 ...
##  $ SubjID    : int  9101 9101 9101 9101 9101 9101 9101 9101 9101 9101 ...
##  $ Difficulty: chr  "Low" "Low" "Low" "Low" ...
##  $ Condition : chr  "Control" "Control" "Control" "Control" ...
##  $ Trial     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Accuracy  : num  0.25 0.25 0.25 0.25 0 0 0 0 0 0 ...
```

```r
MotorLearning$ot1 <- poly(MotorLearning$Trial, 3)[,1]
MotorLearning$ot2 <- poly(MotorLearning$Trial, 3)[,2]
MotorLearning$ot3 <- poly(MotorLearning$Trial, 3)[,3]

m1 <- lmer(Accuracy ~ ot1 + (1 | SubjID), data = MotorLearning, REML = TRUE)
m2 <- lmer(Accuracy ~ ot1 + ot2 + (1 | SubjID), data = MotorLearning, REML = TRUE)
m3 <- lmer(Accuracy ~ ot1 + ot2 + ot3 + (1 | SubjID), data = MotorLearning, REML = TRUE)

AIC(m1, m2, m3)
```

```
##    df      AIC
## m1  4 451.9677
## m2  5 176.4089
## m3  6 157.4618
```

```r
condition_model <- lmer(Accuracy ~ ot1 + ot2 + ot3 + Condition * Difficulty +(1 + ot1 + ot2 + ot3 | Sub
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
summary(condition_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Accuracy ~ ot1 + ot2 + ot3 + Condition * Difficulty + (1 + ot1 +
##     ot2 + ot3 | SubjID) + (1 + ot1 + ot2 + ot3 | SubjID:Condition)
##    Data: MotorLearning
##
## REML criterion at convergence: -628.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2692 -0.5333  0.0755  0.5660  3.0595
##
## Random effects:
##  Groups           Name        Variance Std.Dev. Corr
##  SubjID:Condition (Intercept)  0.01353 0.1163
##                   ot1          4.26519 2.0652    0.59
```

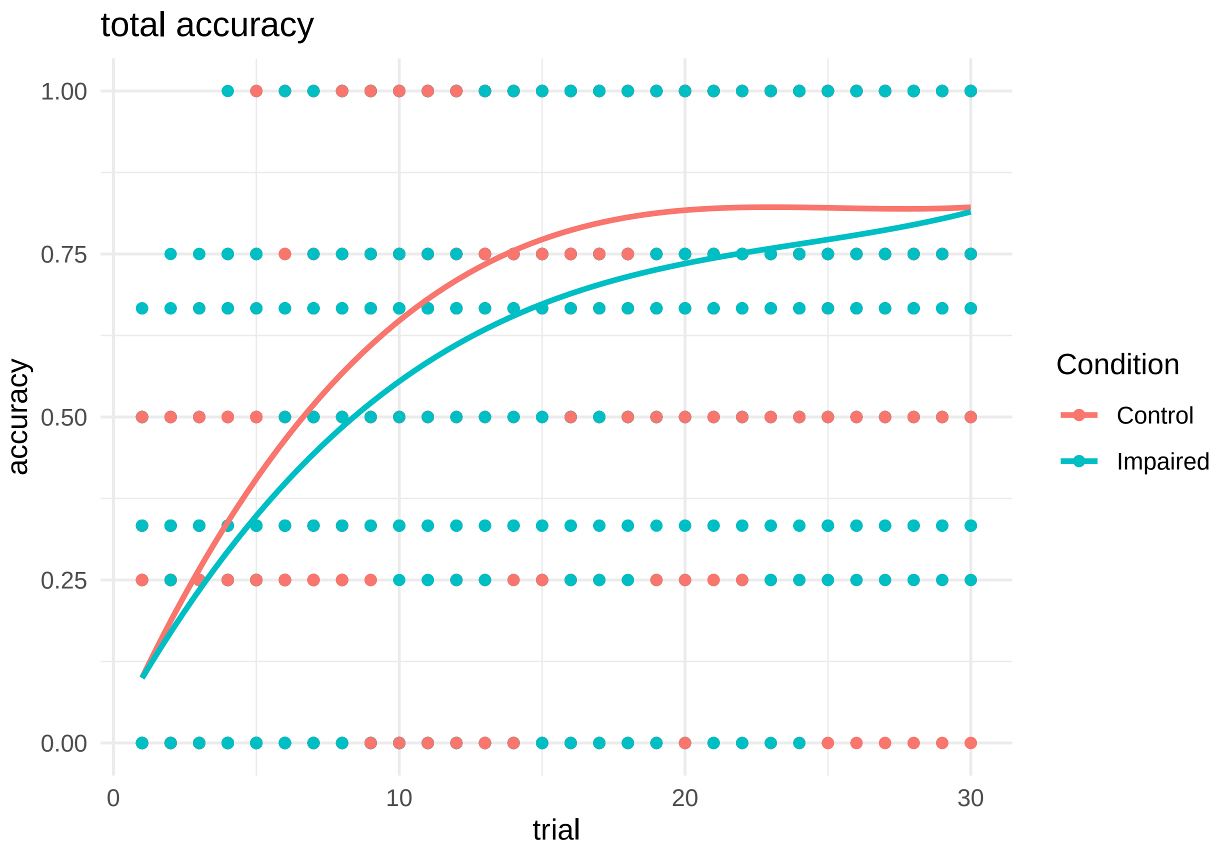
```
##                    ot2          3.66878 1.9154   -0.84 -0.46
##                    ot3          2.27020 1.5067    0.47 -0.40 -0.58
##  SubjID            (Intercept)  0.01636 0.1279
##                    ot1         10.58115 3.2529    0.20
##                    ot2          2.35585 1.5349   -0.75 -0.55
##                    ot3          1.51782 1.2320    0.36 -0.47 -0.44
##  Residual                       0.03989 0.1997
## Number of obs: 2400, groups:  SubjID:Condition, 40; SubjID, 20
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                      0.60481    0.03667  16.495
## ot1                              9.01660    0.82193  10.970
## ot2                             -4.23354    0.49940  -8.477
## ot3                              1.15067    0.41537   2.770
## ConditionImpaired               -0.04726    0.02490  -1.898
## DifficultyLow                    0.09014    0.01153   7.817
## ConditionImpaired:DifficultyLow  0.03042    0.01631   1.865
##
## Correlation of Fixed Effects:
##             (Intr) ot1    ot2    ot3    CndtnI DffclL
## ot1          0.257
## ot2         -0.655 -0.443
## ot3          0.322 -0.369 -0.404
## CondtnImprd -0.340  0.000  0.000  0.000
## DifficltyLw -0.157  0.000  0.000  0.000  0.232
## CndtnImp:DL  0.111  0.000  0.000  0.000 -0.327 -0.707
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

a. Report the results here in APA format. Be sure to describe what kind of pattern of change over time best fits the data (from the unconditional model). What type of relationship there was between the Condition and Difficulty variables, both with each other and the outcome? How does the curve change based on these variables?

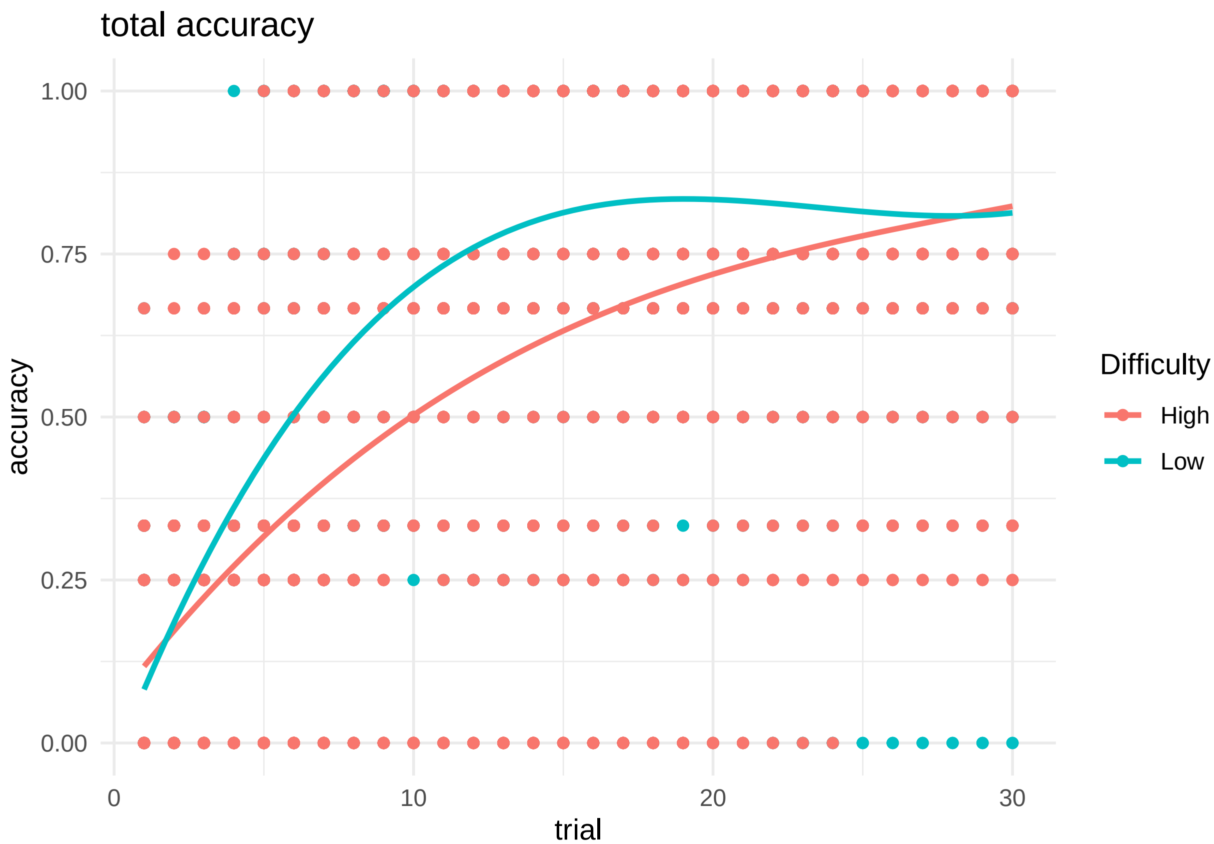
because it is a learning curve of changes over time, analysis used a third-order polynomial equation-the best fit for portraying accuracy that first improves, levels off, and then ultimately stabilizes. Accuracy was higher overall in the control condition than the impaired one. There was a significant interaction wherein the effect of the increased difficulty in the task was even larger when cognitve ability had been impaired. Accuracy curves differed between conditions: in the control condition, accuracy improved consistently, while under the impaired condition, this was much slower and more variable, especially for high-difficulty tasks.

b. Generate a plot that shows the data for each of the conditions and includes the predicted line. And, generate a plot that shows the same, but with the different difficulty levels. Hint: The easiest way is to adapt my ggplot code from the growth curve modeling script.

```r
library(ggplot2)
MotorLearning$Predicted <- predict(condition_model)
ggplot(MotorLearning, aes(x = Trial, y = Accuracy, color = Condition)) +
  geom_point() +
   geom_smooth(method = "lm", formula = y ~ poly(x, 3), se = FALSE)+
  labs(title = "total accuracy",
       x = "trial", y = "accuracy") +
  theme_minimal()
```




```r
ggplot(MotorLearning, aes(x = Trial, y = Accuracy, color = Difficulty)) + geom_point() +
  labs(title = "total accuracy",
       x = "trial", y = "accuracy") +
  theme_minimal()
```


total accuracy
1.00
0.75
0.50
0.25
0.00
accuracy
0
10
20
30
trial
Difficulty
High
Low