Exercises for GLM and GLMM

Exercise 1: determine binary logistic `glmm()` predictions "by hand"

(*requires thinking [and maybe some R-trickery if you want to be lazy]*)

Using the binary logistic GLMM output from slide 37/38 of the lecture on GLMM (the relevant output is repeated below), determine the **predicted *probability* of a correct response per design cell**. There are four such design cells: (1) prime=RB/target=RB; (2) prime=RB/target=LB; (3) prime=LB/target=RB; (4) prime=LB/target=LB. (Notes: Indeed, I want *probabilities* per cell; also consider how the fixed effect predictors ***PT*** and ***TT*** were actually coded in the example, see slide 32/33 of the lecture).

```
     AIC      BIC   logLik deviance df.resid
   891.8    948.9   -433.9    867.8      852

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3626 -0.4189  0.2446  0.4919  2.7978

Random effects:
 Groups Name        Variance Std.Dev.
 subj   (Intercept) 0.4476   0.6691
 subj.1 PT          1.4243   1.1934
 subj.2 TT          9.9545   3.1551
 subj.3 PT:TT       2.5235   1.5886
 item   (Intercept) 0.0000   0.0000
 item.1 PT          0.0000   0.0000
 item.2 TT          0.4498   0.6707
 item.3 PT:TT       0.6734   0.8206
Number of obs: 864, groups:  subj, 36; item, 24

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5572     0.1550   3.595 0.000324 ***
PT           -0.1973     0.2891  -0.682 0.494937
TT            2.1786     0.5939   3.668 0.000244 ***
PT:TT         1.8667     0.5333   3.500 0.000464 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr) PT     TT
PT     0.057
TT     0.041 -0.009
PT:TT  0.047  0.016  0.078
```

Exercise 2: Perform ('maximal') GLMM analyses

(*requires thinking and even more R-trickery*)

The following link takes you to a new dataset (real data, but with 'anonymised', abstract variable names): http://www.psy.gla.ac.uk/~christop/MScStats/2018/HW2data.csv

The data set contains 2433 observations from 43 subjects and 60 items. Some trials (subject-item combinations) are missing due to outlier exclusion.

There are three predictors:
- ***A*** (categorical, with 2 levels *a1* and *a2*) is *between-subjects* but *within-items*
- ***B*** (continuous predictor) is *within-subjects* but *between-items*
- ***C*** (categorical, with 2 levels *c1* and *c2*) is *within-subjects* but *between-items*

The dependent variable (**DV**) is continuous.

(2a) Using mean-centred predictor coding, fit a 3-way ***A × B × C*** GLMM with maximal random effects structure justified by the design to the data (including random correlations). Use ***standard linear*** modelling assumptions for this.

(2b) Fit the model again, but this time assuming a ***Gamma(identity)*** model family.

(2c) Fit the model again, but this time assuming a ***Gamma(log)*** model family.

(2d) Use the ***AIC*** **statistic** to compare the three model fits.

(2e) Which model is the worst, and can you explain why it isn't as good as the other two?

(2f) Look at the model summary for the ***best-fitting*** of the three models. From this summary, identify fixed effects (excluding the intercept) with $p < 0.05$, and report Likelihood Ratio Chi-Squares for those fixed effects.

(2g) look at the model summary for the ***best-fitting*** model again. What appears to be biggest source of random variation in the data?