STAA 554 Homework 2

Contents

Analysis of split plot design

Consider the data: irrigation.csv.

In an agricultural trial the goal was to study the effect of crop variety and crop irrigation technique on yield. Because type of irrigation is challenging to vary on a small scale, an entire irrigation technique is randomized at the field level. Within each field, each variety is randomized to a subplot.

1.) 2pts Table

Create a useful descriptive table below describing yield as a function of irrigation technique and variety. Paste your R code as well as the table below. Make sure the table is polished enough that there are no cell alignment or spacing issues. (many acceptable answers).

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(knitr)

irrigation <- read.csv("Data/irrigation.csv")

desc_table <- irrigation %>%
  group_by(irrigation, variety) %>%
  summarise(Mean_Yield = round(mean(yield, na.rm = TRUE), 2),
            SD_Yield   = round(sd(yield, na.rm = TRUE), 2),
            N          = n()) %>%
  arrange(irrigation, variety)
```

```
## `summarise()` has grouped output by 'irrigation'. You can override using the
## `.groups` argument.
```

```r
kable(desc_table, caption = "Table 1. Yield summarized by Irrigation and Variety")
```

Table 1: Table 1. Yield summarized by Irrigation and Variety

irrigation	variety	Mean_Yield	SD_Yield	N
i1	v1	38.5	4.38	2
i1	v2	39.1	1.70	2
i2	v1	39.7	4.24	2
i2	v2	39.9	2.40	2
i3	v1	39.2	6.22	2
i3	v2	39.6	4.53	2
i4	v1	42.0	3.54	2
i4	v2	43.8	5.37	2

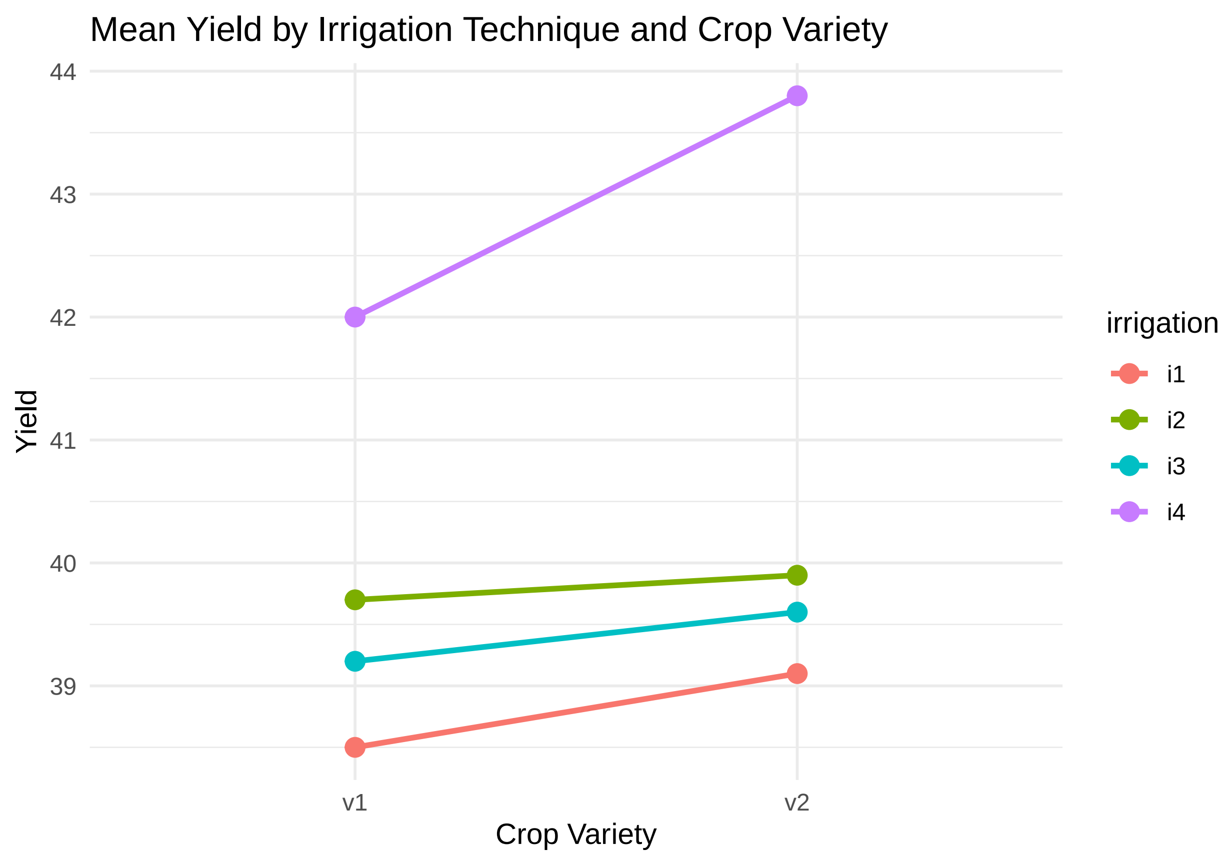
Standard Deviation doesnt really make sense in this case, but there is so little data to analyze that you almost dont even need a summary table

2.) 2 pts Figure

Create a useful figure that summarize yield as a function of irrigation and variety. (many acceptable answers)

```r
library(ggplot2)

ggplot(irrigation, aes(x = variety, y = yield, color = irrigation, group = irrigation)) +
  stat_summary(fun = mean, geom = "point", size = 3) +
  stat_summary(fun = mean, geom = "line", linewidth = 1) +
  labs(title = "Mean Yield by Irrigation Technique and Crop Variety",
       x = "Crop Variety",
       y = "Yield") +
  theme_minimal()
```



3.) 2 pts Model

Fit a model in R that analyzes irrigation type and variety as fixed effects, and field as a random effect. Print a summary of that model below.

```r
library(lme4)
```

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

```r
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step
```

```r
model <- lmer(yield ~ irrigation * variety + (1 | field), data = irrigation)
summary(model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: yield ~ irrigation * variety + (1 | field)
##    Data: irrigation
##
## REML criterion at convergence: 45.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.7448 -0.5509  0.0000  0.5509  0.7448
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  field    (Intercept) 16.200   4.025
##  Residual              2.107   1.452
## Number of obs: 16, groups:  field, 8
##
## Fixed effects:
##                        Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)              38.500      3.026  4.487  12.725 0.000109 ***
## irrigationi2              1.200      4.279  4.487   0.280 0.791591
## irrigationi3              0.700      4.279  4.487   0.164 0.877156
## irrigationi4              3.500      4.279  4.487   0.818 0.454584
## varietyv2                 0.600      1.452  4.000   0.413 0.700582
## irrigationi2:varietyv2   -0.400      2.053  4.000  -0.195 0.855020
## irrigationi3:varietyv2   -0.200      2.053  4.000  -0.097 0.927082
## irrigationi4:varietyv2    1.200      2.053  4.000   0.584 0.590265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) irrgt2 irrgt3 irrgt4 vrtyv2 irr2:2 irr3:2
## irrigation2 -0.707
## irrigation3 -0.707  0.500
## irrigation4 -0.707  0.500  0.500
## varietyv2   -0.240  0.170  0.170  0.170
## irrgtn2:vr2  0.170 -0.240 -0.120 -0.120 -0.707
## irrgtn3:vr2  0.170 -0.120 -0.240 -0.120 -0.707  0.500
## irrgtn4:vr2  0.170 -0.120 -0.120 -0.240 -0.707  0.500  0.500
```

4.) 2 pts Design

Suppose you are concerned that effect of variety on yield may depend on field. How would this design need to be modified to model a field and variety interaction effect? In other words, what is preventing you from estimating that effect from this design? (Applies to fixed effect model as well.)

In the current design each field has one observation per variety. As such, the effect is confounded with the field - level error term. This means that we cant estimate the true field by variety interaction as there are no replicates.

5a.) 1pt Test

Based on the mixed model from part "3", use anova() to perform an F test for the statistical significance of irrigationXvariety interaction. State your conclusion.

```
anova(model)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                     Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## irrigation          2.4545 0.81818     3     4  0.3882 0.7685
## variety             2.2500 2.25000     1     4  1.0676 0.3599
## irrigation:variety 1.5500 0.51667     3     4  0.2452 0.8612
```

There is no statistically signifficant interaction between irrigation and variety as the P value is .8

5b.) 1pt Test

Using anova() perform an F test for the statistical significance of variety.

With a P value of .35, there is no signifficant affect of variety on yield.

6) 2 pts Inference

How does each irrigation technique compare to the others for a particular variety? Use the emmeans() command to address this question. Paste your output below, be sure to comment.

```
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```
emm <- emmeans(model, ~ irrigation | variety)
summary(emm)
```

```
## variety = v1:
##  irrigation emmean   SE   df lower.CL upper.CL
##  i1           38.5 3.03 4.49     30.4     46.6
##  i2           39.7 3.03 4.49     31.6     47.8
```

```
##  i3           39.2 3.03 4.49     31.1     47.3
##  i4           42.0 3.03 4.49     33.9     50.1
##
## variety = v2:
##  irrigation emmean   SE   df lower.CL upper.CL
##  i1           39.1 3.03 4.49     31.0     47.2
##  i2           39.9 3.03 4.49     31.8     48.0
##  i3           39.6 3.03 4.49     31.5     47.7
##  i4           43.8 3.03 4.49     35.7     51.9
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
```

```r
pairs(emm)
```

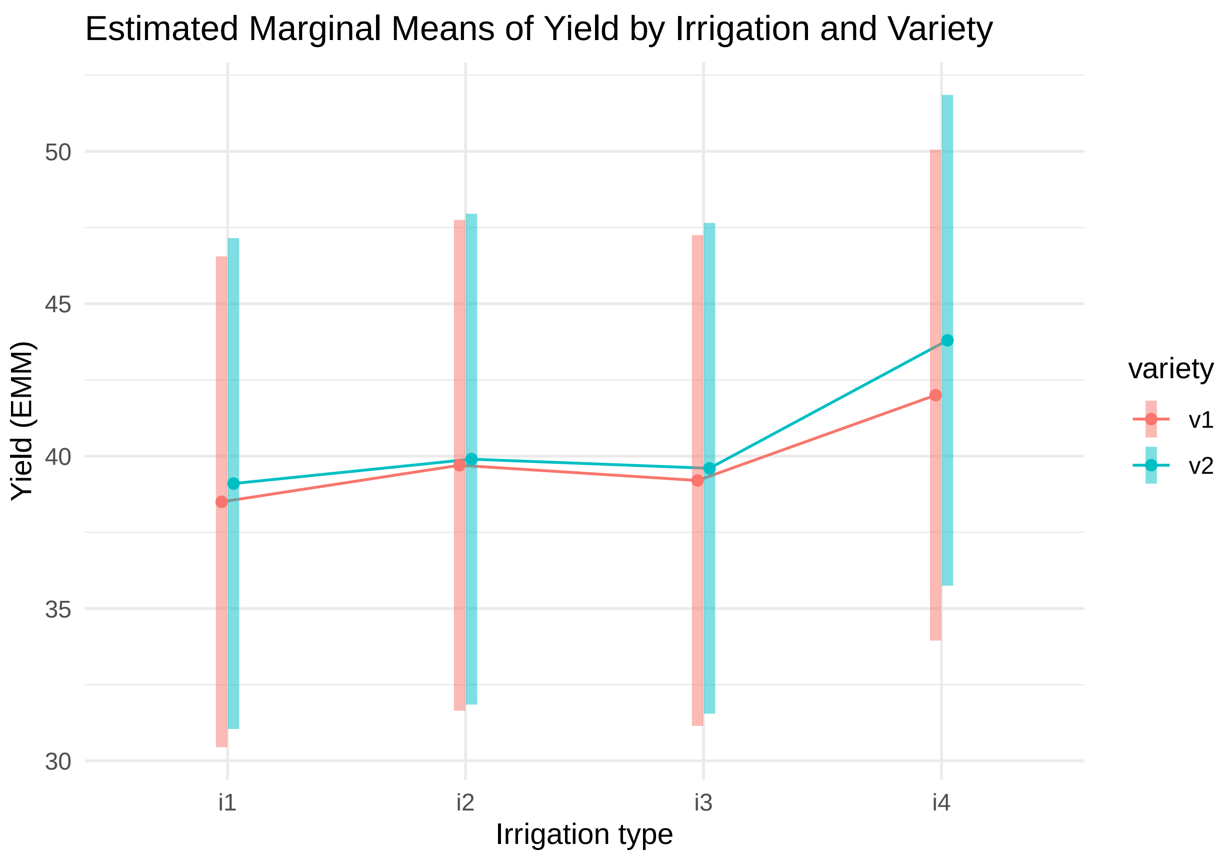
```
## variety = v1:
##  contrast estimate   SE   df t.ratio p.value
##  i1 - i2      -1.2 4.28 4.49  -0.280  0.9913
##  i1 - i3      -0.7 4.28 4.49  -0.164  0.9982
##  i1 - i4      -3.5 4.28 4.49  -0.818  0.8440
##  i2 - i3       0.5 4.28 4.49   0.117  0.9993
##  i2 - i4      -2.3 4.28 4.49  -0.538  0.9457
##  i3 - i4      -2.8 4.28 4.49  -0.654  0.9093
##
## variety = v2:
##  contrast estimate   SE   df t.ratio p.value
##  i1 - i2      -0.8 4.28 4.49  -0.187  0.9973
##  i1 - i3      -0.5 4.28 4.49  -0.117  0.9993
##  i1 - i4      -4.7 4.28 4.49  -1.098  0.7075
##  i2 - i3       0.3 4.28 4.49   0.070  0.9999
##  i2 - i4      -3.9 4.28 4.49  -0.911  0.8009
##  i3 - i4      -4.2 4.28 4.49  -0.982  0.7667
##
## Degrees-of-freedom method: kenward-roger
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Emmeans also finds no signifficant relaionships between any of the irrigation techniques

7) 2 pts Figure

Provide a figure summarizing the findings in part 6.

```r
emmip(model, variety ~ irrigation, CIs = TRUE) +
  labs(title = "Estimated Marginal Means of Yield by Irrigation and Variety",
       x = "Irrigation type",
       y = "Yield (EMM)") +
  theme_minimal()
```



Estimated Marginal Means of Yield by Irrigation and Variety
Yield (EMM)
50
45
40
35
30
i1
i2
i3
i4
Irrigation type
variety
v1
v2

Analysis of Observational Data

This dataset is a slightly modified version of the dataset used for a special issue of Leadership Quarterly (Vol. 12, 2002), edited by Paul Bliese, Ronald Halverson, and Chet Schriesheim. It contains the items for three psychological scales: Hostility (HOSTILE), Task Significance (TSIG), and Leadership Climate (LEAD). There are 2,042 observations clustered within 49 groups, which represent Army companies. The scales were measured at the individual level and then aggregated at the company level. Data are public domain and are used with the kind permission of Paul Bliese (Walter Reid Army Institute of Research). We focus on a subset of the variables in the data:

```
COMPID  numeric  Army Company Identifying Variable

SUB  numeric  Subject Number

LEAD  numeric  Leadership Climate Scale Score

TSIG  numeric  Task Significance Scale Score

HOSTILE  numeric  Hostility Scale Score

GLEAD  numeric  Leadership Climate Score Aggregated By Company

GTSIG  numeric  Task Significance Score Aggregated By Company

GHOSTILE  numeric  Hostility Score Aggregated By Company
```

Research Questions:

- What is the individual-level (TSIG) and company-level (GTSIG) perception of task significance on the mean individual-level feelings of hostility (HOSTILE)? (adjusting for the clustering of hostility scale scores by army company)
- What is the effect of soldier-level perceived leadership climate (LEAD) on HOSTILE, after controlling for other effects in the model.

8. 2pts Data Summaries

Provide tables for the level 1 and level 2 variables included in the description above. You should perform your own exploratory data analysis as well, but we will not focus on that for grading here.

```r
IQ <- read.csv("Data/lq2002.csv")

level1_vars <- IQ[, c("SUB", "LEAD", "TSIG", "HOSTILE")]
level2_vars <- IQ[, c("COMPID", "GLEAD", "GTSIG", "GHOSTILE")]

kable(summary(level1_vars), caption = "Summary of Level 1 Variables")
```

Table 2: Summary of Level 1 Variables

SUB	LEAD	TSIG	HOSTILE
Min. : 1.0	Min. :1.000	Min. :1.000	Min. :0.000

SUB	LEAD	TSIG	HOSTILE
1st Qu.: 511.2	1st Qu.:2.455	1st Qu.:2.333	1st Qu.:0.200
Median :1021.5	Median :3.000	Median :3.333	Median :0.600
Mean :1021.5	Mean :3.001	Mean :3.133	Mean :0.947
3rd Qu.:1531.8	3rd Qu.:3.636	3rd Qu.:4.000	3rd Qu.:1.600
Max. :2042.0	Max. :5.000	Max. :5.000	Max. :4.000

```
kable(summary(level2_vars), caption = "Summary of Level 2 Variables")
```

Table 3: Summary of Level 2 Variables

COMPID	GLEAD	GTSIG	GHOSTILE
Min. : 2.00	Min. :2.379	Min. :2.541	Min. :0.2154
1st Qu.:18.00	1st Qu.:2.811	1st Qu.:2.868	1st Qu.:0.7200
Median :29.00	Median :2.984	Median :3.124	Median :0.9077
Mean :28.73	Mean :3.001	Mean :3.133	Mean :0.9470
3rd Qu.:41.00	3rd Qu.:3.176	3rd Qu.:3.394	3rd Qu.:1.1741
Max. :58.00	Max. :3.739	Max. :3.778	Max. :1.5206

9. 2pts ICC

Calculate the raw ICC. What does this tell you?

```
IQ_null <- lmer(HOSTILE ~ 1 + (1 | COMPID), data = IQ, REML = TRUE)

var_COMPID <- as.numeric(VarCorr(IQ_null)$COMPID)
var_residual <- sigma(IQ_null)^2

icc <- var_COMPID / (var_COMPID + var_residual)
print(icc)
```

```
## [1] 0.05528403
```

This tells me that 5% of the variation in Hostility is explained by differences between companies

10. 2pts Model Specification

Based on the R-code for the model give below, right out the Level 1 and Level 2 models. (Define what your subscripts index, e.g. i: individual, j: company)
`lmer(HOSTILE ~ TSIG + GTSIG + (1|COMPID), data = df)`

Level-1 Model (Within-Company):

$$\text{HOSTILE}_{ij} = \beta_{0j} + \beta_1 \, \text{TSIG}_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2)$$

Level-2 Model (Between-Company):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \, \text{GTSIG}_j + u_{0j}, \quad u_{0j} \sim N(0, \tau^2)$$

- β_{0j} is the intercept for company j, which varies across companies,
- β_1 is the fixed effect of the individual-level task significance (TSIG) on hostility,
- γ_{00} is the overall intercept across all companies,
- γ_{01} is the effect of the company-level task significance (GTSIG) on the company intercept,
- r_{ij} represents the residual error at the individual level, and
- u_{0j} is the random effect for company j.

11. 2pt Variance Components

Fit the model below with and without GTSIG. How does removing this term affect the variance components? Explain why this makes sense.
`mod1 = lmer(HOSTILE ~ TSIG + GTSIG + (1|COMPID), data = df)`

```r
mod1 <- lmer(HOSTILE ~ TSIG + GTSIG + (1 | COMPID), data = IQ, REML = TRUE)
mod_noGTSIG <- lmer(HOSTILE ~ TSIG + (1 | COMPID), data = IQ, REML = TRUE)
summary(mod1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: HOSTILE ~ TSIG + GTSIG + (1 | COMPID)
##    Data: IQ
##
## REML criterion at convergence: 5668
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9304 -0.7026 -0.2675  0.4856  3.4294
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  COMPID   (Intercept) 0.02045  0.1430
##  Residual             0.92124  0.9598
## Number of obs: 2042, groups:  COMPID, 49
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)    2.73164    0.29853   36.01164   9.150 6.28e-11 ***
## TSIG          -0.32107    0.02196 1991.20232 -14.620  < 2e-16 ***
## GTSIG         -0.25226    0.09667   41.48247  -2.609   0.0126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) TSIG
## TSIG   0.000
## GTSIG -0.969 -0.227
```

```r
summary(mod_noGTSIG)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: HOSTILE ~ TSIG + (1 | COMPID)
##    Data: IQ
##
## REML criterion at convergence: 5671.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.9626 -0.7132 -0.2692  0.4721  3.4835
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  COMPID   (Intercept) 0.02947  0.1717
##  Residual             0.91978  0.9591
## Number of obs: 2042, groups:  COMPID, 49
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)    1.96414    0.07599  548.78416   25.85   <2e-16 ***
## TSIG          -0.33149    0.02147 1994.85950  -15.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## TSIG -0.892
```

Including GTSIG explains some of the hostile variability that is attribuatable to company differences. When we omit it, this variability is abosrbed by the random effect. This increases the variance of our random intercept

12. 2pt Information Criteria

While REML is the recommended (and default) method for fitting models. If we use ML we can then use information Criteria such as AIC, BIC, and DIC. Fit the below models and choose one based on an information criteria. Justify why you chose this model.

```
mod1 = lmer(HOSTILE ~ TSIG + GTSIG + (1|COMPID), data = df, REML = FALSE)
mod2 = lmer(HOSTILE ~ TSIG + GTSIG + LEAD +(1|COMPID), data = df, REML= FALSE)
```

```r
mod1_ML <- lmer(HOSTILE ~ TSIG + GTSIG + (1 | COMPID), data = IQ, REML = FALSE)
mod2_ML <- lmer(HOSTILE ~ TSIG + GTSIG + LEAD + (1 | COMPID), data = IQ, REML = FALSE)
AIC(mod1_ML, mod2_ML)
```

```
##         df      AIC
## mod1_ML  5 5664.119
## mod2_ML  6 5533.666
```

```r
BIC(mod1_ML, mod2_ML)
```

```
##         df      BIC
## mod1_ML  5 5692.227
## mod2_ML  6 5567.396
```

```r
summary(mod2_ML)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: HOSTILE ~ TSIG + GTSIG + LEAD + (1 | COMPID)
##    Data: IQ
##
##      AIC      BIC   logLik deviance df.resid
##   5533.7   5567.4  -2760.8   5521.7     2036
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3166 -0.6960 -0.2211  0.5170  3.4998
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  COMPID   (Intercept) 0.009905 0.09952
##  Residual             0.867050 0.93116
## Number of obs: 2042, groups:  COMPID, 49
##
## Fixed effects:
##              Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)   3.40718    0.25785   42.52614  13.214  < 2e-16 ***
## TSIG         -0.17234    0.02479 2041.22422  -6.953 4.79e-12 ***
## GTSIG        -0.27729    0.08258   46.65424  -3.358  0.00157 **
## LEAD         -0.35177    0.02996 1824.44441 -11.741  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) TSIG   GTSIG
## TSIG   0.105
## GTSIG -0.942 -0.227
## LEAD  -0.205 -0.511  0.010
```

Model 2 is preferred by both AIC and BIC measures. this implies that adding the LEAD value improves model fit despite the extra parameter.

13. 2pt Inference

Use the models fit above to answer the research question:

- What is the individual-level (TSIG) and company-level (GTSIG) perception of task significance on the mean individual-level feelings of hostility (HOSTILE)? (adjusting for the clustering of hostility scale scores by army company)

Justify your conclusion by citing the relevant test statistic.
Hint: you may want to load the lmerTest package.

The model shows that both levels are significant predictors of hostility. a one unit increase in individual task significance is associated with a decrease of about .32 units and a one unit increase of task signifficance at the company level is associated with a decrease of around .25 units. This relationship is significant with p values of 2e-16 and .0126 respectively.

14. 2pt Inference

Use the models fit above to answer the research question:

- What is the effect of soldier-level perceived leadership climate (LEAD) on HOSTILE, after controlling for other effects in the model.

Justify your conclusion by citing the relevant test statistic.

The coefficient for LEAD is −0.351 and is signifficant with p value of 2e-16. This shows that, after accounting for task significance at both levels, a one-unit increase in perceived leadership climate is associated with a 0.35 unit decrease in HOSTILE. This demonstrates that leadership climate significantly reduces individual hostility.