

Homework 5

Psych 5068

Emorie Beck

March 5, 2018

Contents

Packages	2
Data	2
Question 1.	2
Question 2	4
Question 3	5
Part A	5
Part B	6
Part C	7
Part D	7
Question 4	8
Part A	8
Part B	9
Part C	9
Part D	10
Question 5	10
Part A	10
Part B	10
Part C	11
Part D	12
Part E	12
Part F	12
Question 6	13
Question 7	14
Question 8	15
Part A	16
Part B	17
Part C	18
Question 9	19
Part A	19
Part B	20
Question 10	21
Part A	22
Part B	22

data (popularity.csv) to explore methods for examining the adequacy of hierarchical linear models. Begin with the unconditional model:

```
# Workspace
```

Packages

```
library(psych)
library(lme4)
library(knitr)
library(qqplotr)
library(influence.ME)
library(HLMdiag)
library(kableExtra)
library(plyr)
library(tidyverse)
```

Data

```
data_url <- "https://raw.githubusercontent.com/emoriebeck/homeworks/master/homework5/popularity(6).csv"
dat <- read.csv(url(data_url)) %>% tbl_df %>%
  mutate(sex12 = sex,
         sex = factor(sex, levels = 0:1, labels = c("Male", "Female")))
```

Question 1.

Test the homogeneity of Level 1 residual variances (σ^2) assumption by comparing a model that estimates a single Level 1 variance (the default, call it Pop_Fit_1) to a model that estimates a separate variance for each classroom (call it Pop_Fit_2).

Reminder: This needs to be done with the nlme package, which allows specifying separate variances for each Level 2 unit.

Level 1:

$$popular_{ij} = \beta_{0j} + r_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

```
source("https://raw.githubusercontent.com/emoriebeck/homeworks/master/table_fun.R")
library(nlme)
Pop_Fit_1 <- lme(popular ~ 1, random = ~1 | class, data = dat)
Pop_Fit_2 <- lme(popular ~ 1, random = ~1 | class, data = dat, varIdent(form = ~ 1 | class))

Pop_Fit_1 %>% summary

## Linear mixed-effects model fit by REML
## Data: dat
##      AIC      BIC    logLik
## 6336.51 6353.311 -3165.255
##
## Random effects:
```

```
## Formula: ~1 | class
##      (Intercept) Residual
## StdDev:   0.8379169 1.105348
##
## Fixed effects: popular ~ 1
##              Value Std.Error   DF  t-value p-value
## (Intercept) 5.07786 0.08739443 1900 58.10279      0
##
## Standardized Within-Group Residuals:
##      Min      Q1      Med      Q3      Max
## -3.565536840 -0.697542781  0.001956985  0.675810799  3.317504350
##
## Number of Observations: 2000
## Number of Groups: 100
```

Pop_Fit_2 %>% summary

```
## Linear mixed-effects model fit by REML
## Data: dat
##      AIC      BIC    logLik
## 6405.673 6976.914 -3100.837
##
## Random effects:
## Formula: ~1 | class
##      (Intercept) Residual
## StdDev:   0.8379648 0.9521529
##
## Variance function:
## Structure: Different standard deviations per stratum
## Formula: ~1 | class
## Parameter estimates:
##      1      2      3      4      5      6      7
## 1.0000000 1.0499363 1.2207029 0.8541240 0.9973815 1.0157321 1.1445444
##      8      9     10     11     12     13     14
## 1.2170772 1.3346129 0.8740021 0.9586056 1.0060780 0.8900662 1.3095554
##     15     16     17     18     19     20     21
## 1.2881853 1.2804569 1.2384373 1.0313981 1.1937506 1.4532934 0.9313162
##     22     23     24     25     26     27     28
## 0.8019151 1.1213246 1.3330274 1.1738459 1.1501272 1.1511586 0.8015011
##     29     30     31     32     33     34     35
## 0.9217862 1.2674700 1.8038285 1.2396037 1.4369198 1.3081363 0.6554007
##     36     37     38     39     40     41     42
## 1.0910736 1.2617634 1.0847395 0.7528253 1.4065366 0.8272193 1.2179219
##     43     44     45     46     47     48     49
## 0.8003435 1.0748452 1.2180299 1.3702985 0.9218534 1.6334850 0.9792693
##     50     51     52     53     54     55     56
## 0.8405058 1.0975264 1.1401357 0.9160148 1.0518165 1.0291801 1.2322453
##     57     58     59     60     61     62     63
## 1.5467456 1.2170930 1.1781249 1.5876935 1.0859182 1.1135284 1.2075701
##     64     65     66     67     68     69     70
## 1.3015184 1.1231329 1.2757089 0.9583205 0.9910443 1.0042414 1.1896880
##     71     72     73     74     75     76     77
## 1.3150873 1.0775448 1.0769280 1.0474515 0.9294291 0.8908559 1.3637457
##     78     79     80     81     82     83     84
## 1.0494309 1.3123283 1.1232540 1.0290107 1.3425965 0.8867385 1.0345774
```

```
##          85          86          87          88          89          90          91
## 0.9229351 1.2415380 1.5730716 1.2662948 1.3487742 0.9957063 1.1889781
##          92          93          94          95          96          97          98
## 0.9918226 1.1231373 1.0658333 1.0113186 1.2665757 1.2133722 0.9214493
##          99          100
## 1.2959411 1.8378429
## Fixed effects: popular ~ 1
##              Value Std.Error   DF t-value p-value
## (Intercept) 5.085095 0.08733616 1900 58.2244      0
##
## Standardized Within-Group Residuals:
##          Min          Q1          Med          Q3          Max
## -4.222647455 -0.817049909 -0.009953112  0.767681205  3.845590285
##
## Number of Observations: 2000
## Number of Groups: 100
anova(Pop_Fit_1,Pop_Fit_2)
```

```
##          Model df          AIC          BIC    logLik    Test  L.Ratio p-value
## Pop_Fit_1      1    3 6336.510 6353.311 -3165.255
## Pop_Fit_2      2 102 6405.673 6976.914 -3100.837 1 vs 2 128.8365  0.0236
```

Thus, we do not meet the homogeneity of variance assumption. The model that estimates separate variances is slightly better than one that does not.

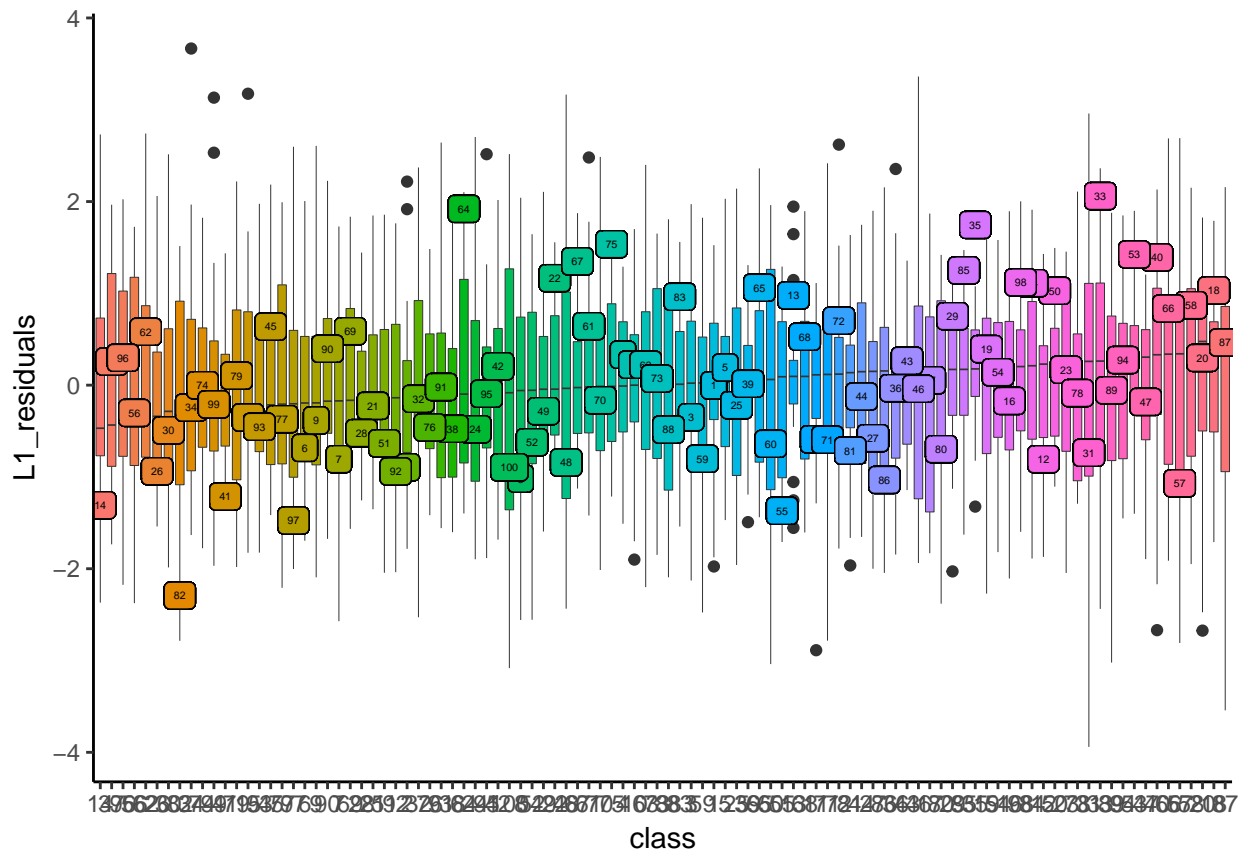
Question 2

Add the Level 1 residuals (and fitted values) from Pop_Fit_1 to the popularity data file (name them L1_residuals and L1_fitted, respectively) and produce a boxplot figure showing the residual distributions by classroom (class).

```
dat <- broom::augment(Pop_Fit_1) %>% tbl_df %>%
  mutate(class = as.character(class)) %>%
  full_join(
    rane(Pop_Fit_1) %>% data.frame %>% setNames("L2_residuals") %>%
    mutate(class = rownames(.)) %>% tbl_df
  ) %>% rename(L1_residuals = .resid)

orders <- dat %>%
  group_by(class) %>%
  summarize(median = median(L1_residuals, na.rm = T)) %>%
  arrange(median)

dat %>%
  mutate(class = factor(class, levels = orders$class)) %>%
  ggplot(aes(x = class, y = L1_residuals, fill = class)) +
  geom_boxplot(size = .15) +
  geom_label(data = dat, aes(y = L2_residuals, label = class), size = 1.5) +
  theme_classic() +
  theme(legend.position = "none")
```



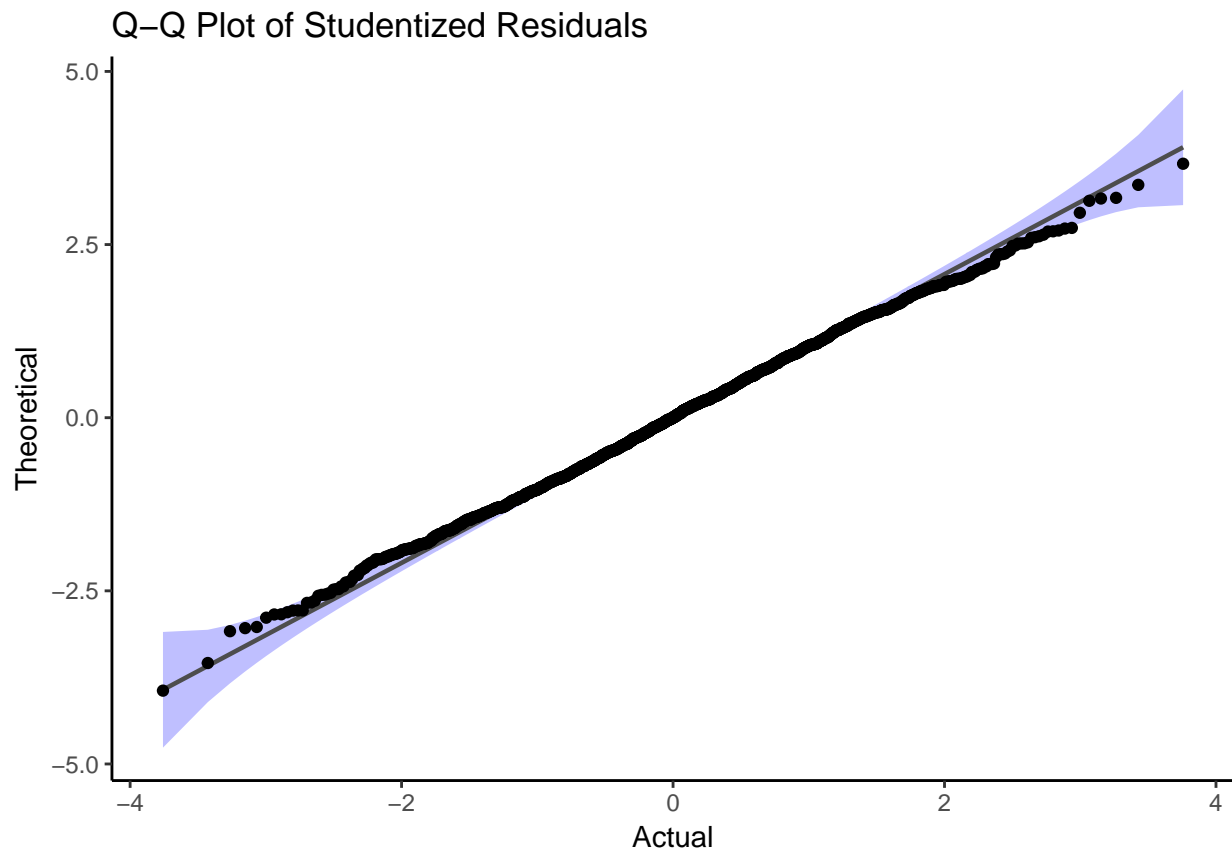
Question 3

Examine the Level 1 residuals:

Part A

Construct a Q-Q plot of the Level 1 residuals.

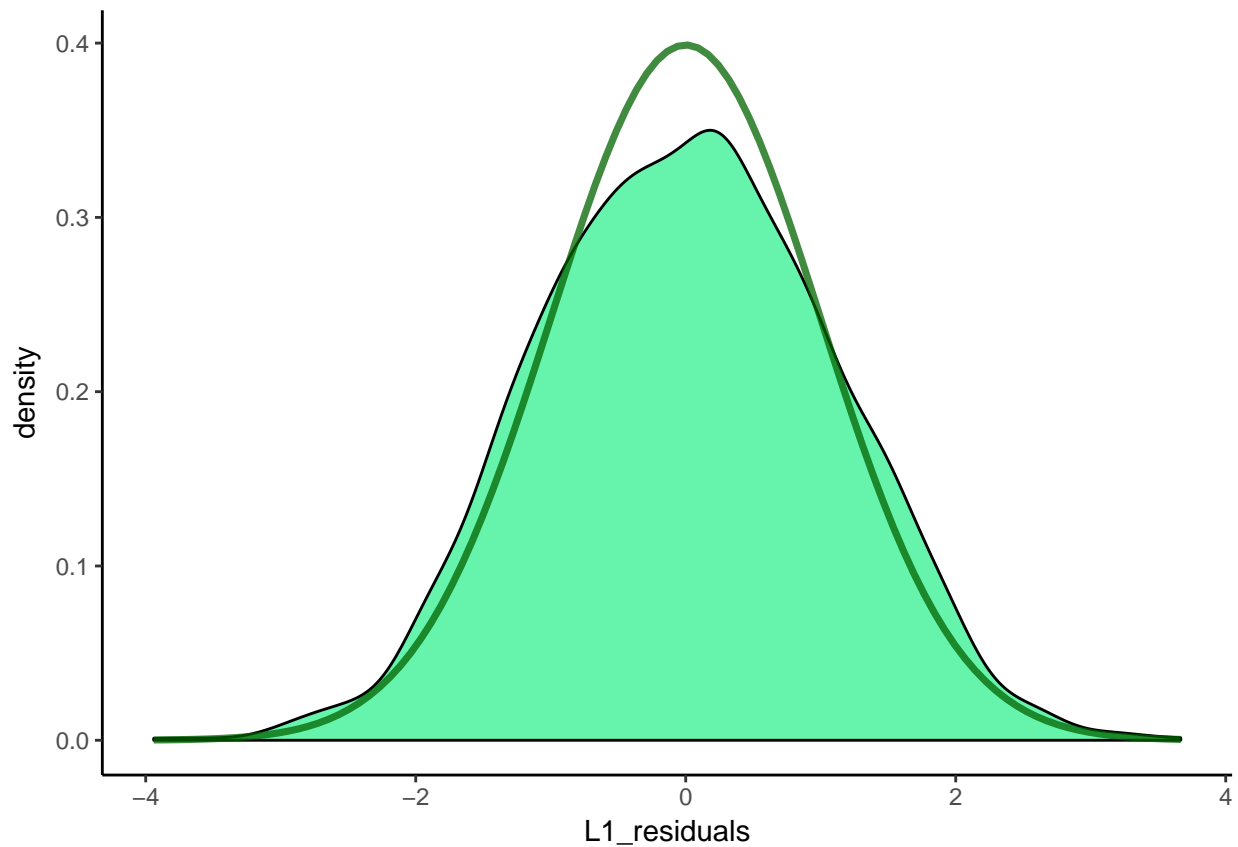
```
dat %>%
  ggplot(aes(sample=L1_residuals)) +
  stat_qq_band(fill = "blue", alpha = .25) +
  stat_qq_line() +
  stat_qq_point() +
  labs(x = "Actual", y = "Theoretical",
       title = "Q-Q Plot of Studentized Residuals") +
  theme_classic()
```



Part B

Construct a histogram of the Level 1 residuals with a normal distribution overlay.

```
dat %>%  
  ggplot(aes(x = L1_residuals)) +  
  geom_density(fill = "springgreen2", alpha = .6) +  
  stat_function(fun = dnorm, size = 1.25, color = "darkgreen", alpha = .75,  
               args = list(mean = 0, sd = 1)) +  
  theme_classic()
```



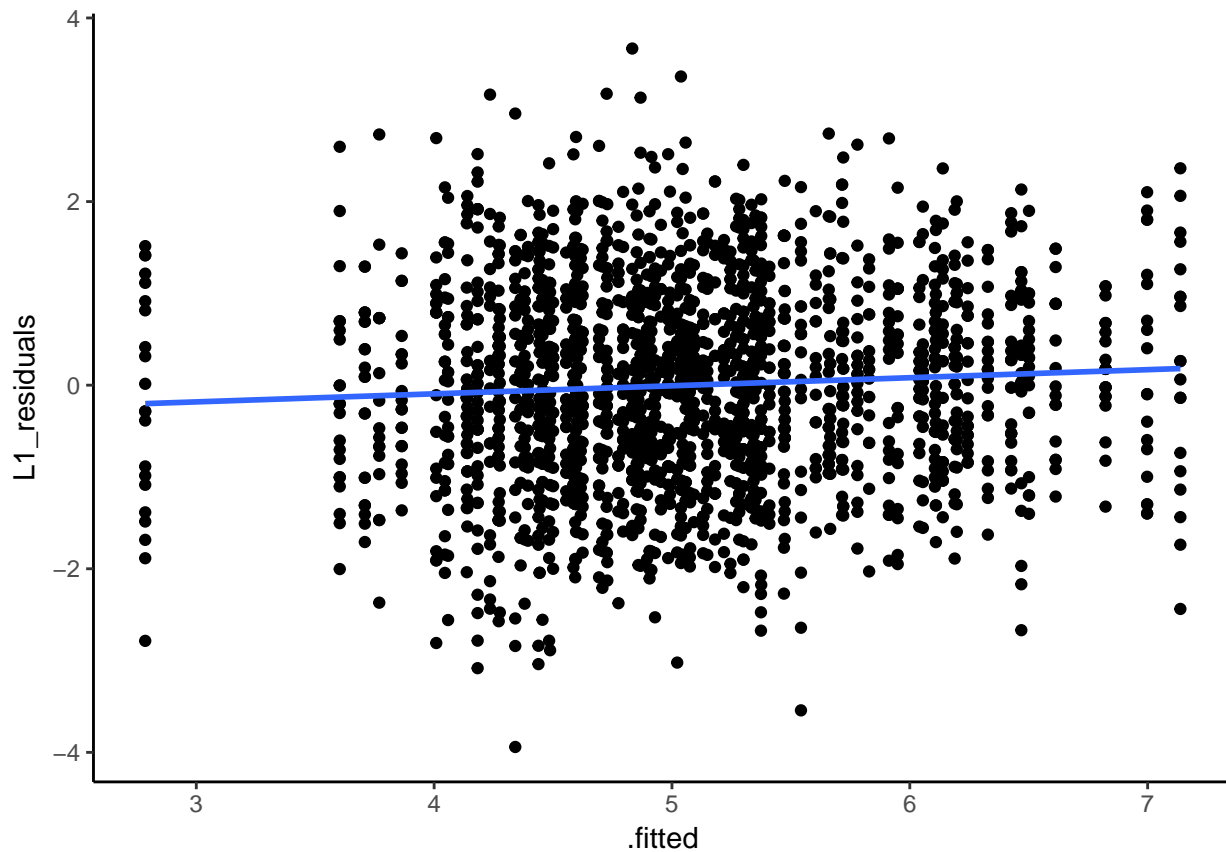
Part C

Are the Level 1 residuals normally distributed? Yes, the level 1 residuals are normally distributed.

Part D

Construct a scatterplot of the Level 1 residuals against the Level 1 fitted values. Comment on the assumption of homoscedasticity and what that means for this unconditional model.

```
dat %>%  
  ggplot(aes(x = .fitted, y = L1_residuals)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = F) +  
  theme_classic()
```



The plot of Level 1 residuals v. Level 1 fitted values suggests that we meet the assumption of homoscedasticity. There is a very weak positive relationship, but not strong enough to worry.

Question 4

Now determine if Level 1 predictors should be added to the model.

Part A

Correlate the Level 1 residuals with extraversion. Is there evidence that this predictor should be included?

```
r <- dat %>% select(extrav, sex12, .fitted, L1_residuals) %>% cor

r[upper.tri(r, diag = T)] <- NA
r <- r %>% data.frame() %>% mutate(v1 = rownames(.)) %>% select(v1, everything())

options(knitr.kable.NA = '')
r %>%
  kable(., "latex", digits = 2, booktabs = T,
        col.names = c("", "extrav", "sex", "fitted", "L1 resid"))
```

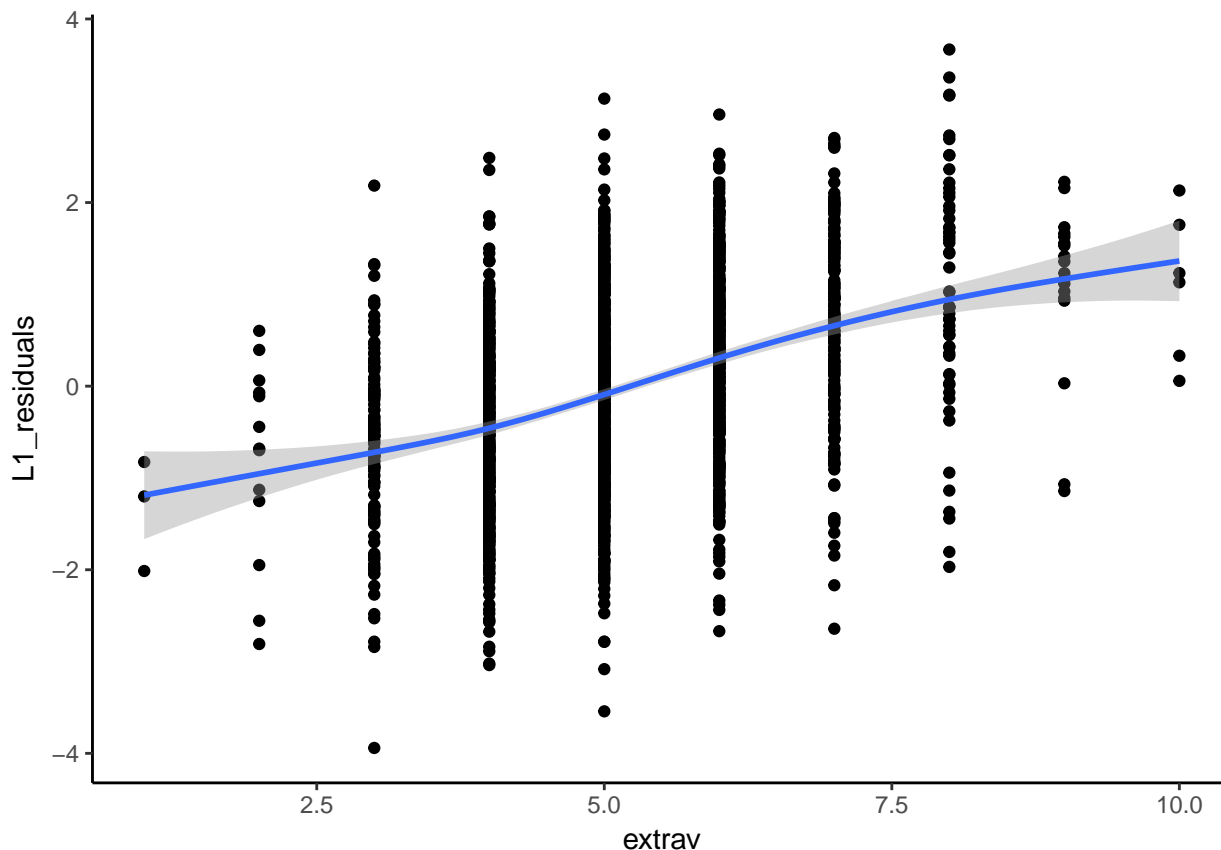

	extrav	sex	fitted	L1 resid
extrav				
sex12	0.09			
.fitted	-0.01	0.26		
L1_residuals	0.41	0.54	0.06	

Extraversion is moderately correlated with the level 1 residuals ($r = 0.41$), suggesting that it should be included in the model.

Part B

Create a scatterplot showing the relationship between the Level 1 residuals and extraversion. Does there appear to be any need to model nonlinearity?

```
dat %>%
  ggplot(aes(x = extrav, y = L1_residuals)) +
  geom_point() +
  geom_smooth() +
  theme_classic()
```



No, there is possibly a very small, non-linear effect, but the degree of nonlinearity is very small.

Part C

Correlate the Level 1 residuals with student sex. Is there evidence that this predictor should be included? Sex is moderately to strongly correlated with the level 1 residuals ($r = 0.54$), suggesting that it should be

included in the model.

Part D

Should both predictors be included in the model? That is, do they appear to be unique predictors?

Extraversion and sex are nearly uncorrelated ($r = 0.09$), suggesting that they do appear to be unique predictors.

Question 5

Examine the Level 2 residuals:

Part A

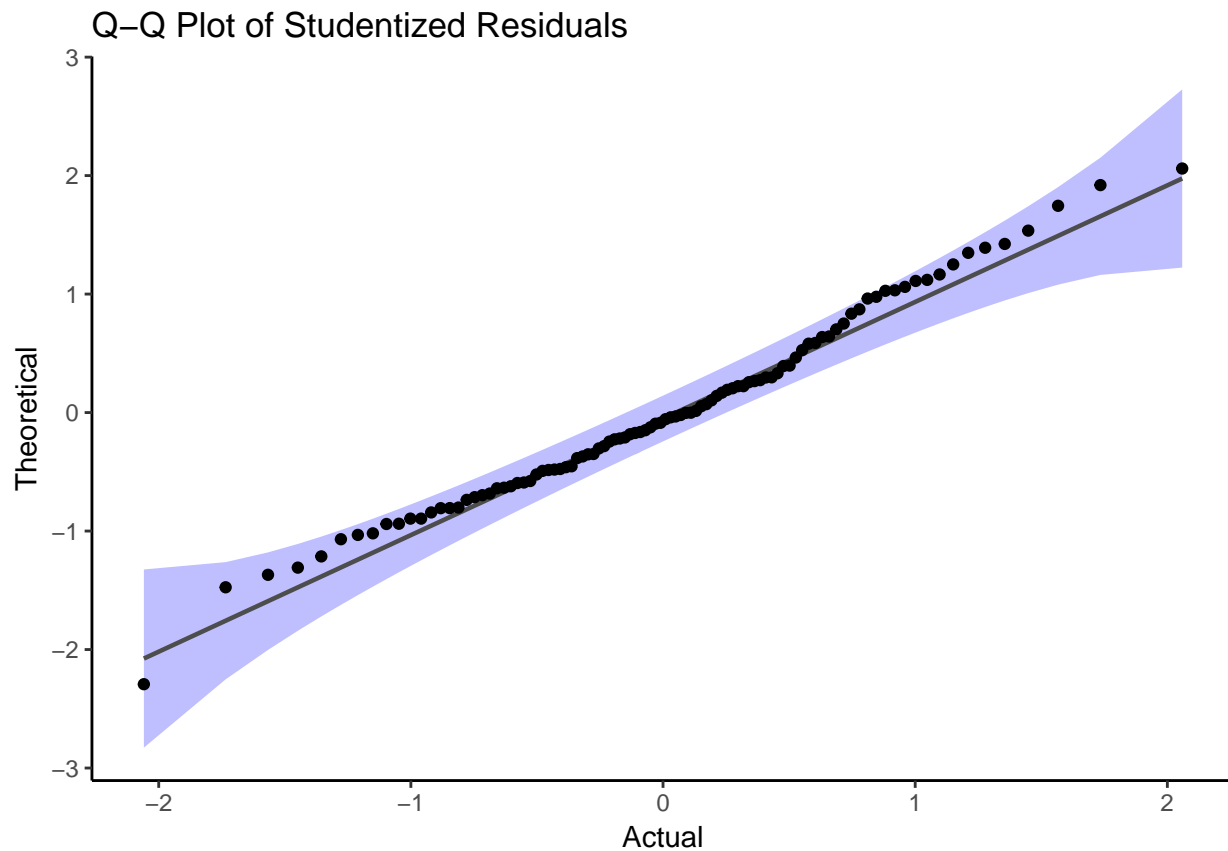
Create a classroom level data frame (call it `Class_Data`) that contains the Level 2 residuals (only intercept residuals are available so far; name it `R_Intercept`) and the grand-mean centered classroom means for extraversion (name it `Mean_E_GMC`).

```
Class_Data <- dat %>%  
  mutate(gmc = mean(extrav, na.rm = T)) %>%  
  group_by(class) %>%  
  summarise(Mean_E_GMC = mean(extrav, na.rm = T)/unique(gmc)) %>%  
  full_join(unique(dat %>% select(class, L2_residuals)))
```

Part B

Construct a Q-Q plot of the Level 2 residuals

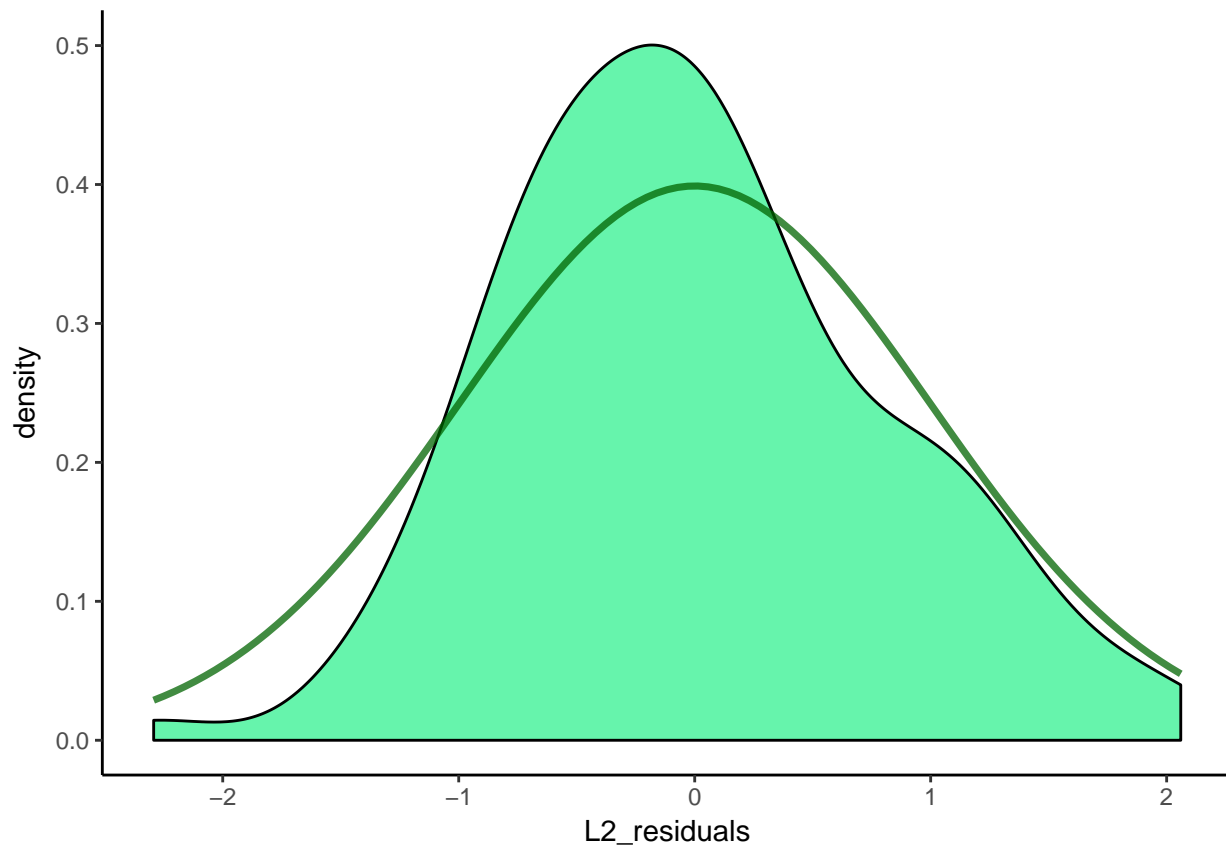
```
Class_Data %>%  
  ggplot(aes(sample=L2_residuals)) +  
  stat_qq_band(fill = "blue", alpha = .25) +  
  stat_qq_line() +  
  stat_qq_point() +  
  labs(x = "Actual", y = "Theoretical",  
        title = "Q-Q Plot of Studentized Residuals") +  
  theme_classic()
```



Part C

Construct a histogram of the Level 2 residuals with a normal distribution overlay.

```
Class_Data %>%  
  ggplot(aes(x = L2_residuals)) +  
  geom_density(fill = "springgreen2", alpha = .6) +  
  stat_function(fun = dnorm, size = 1.25, color = "darkgreen", alpha = .75,  
               args = list(mean = 0, sd = 1)) +  
  theme_classic()
```



Part D

Are the Level 2 residuals normally distributed? The Level 2 residuals do not appear to be normally distributed.

Part E

Correlate the Level 2 residuals with classroom mean extraversion. Is there evidence that this predictor should be included in the Level 2 model?

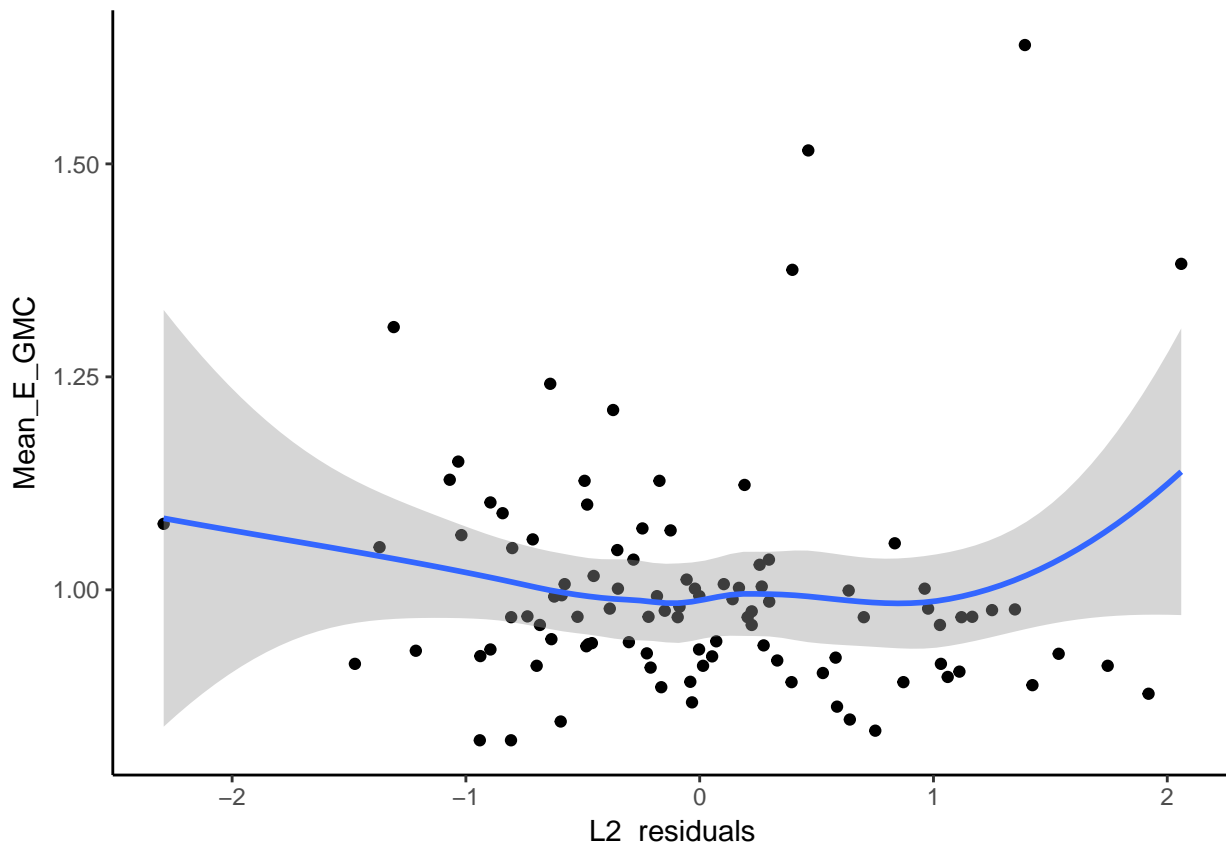
```
r2 <- Class_Data %>% summarize(r = cor(Mean_E_GMC, L2_residuals))
```

Classroom mean extraversion is almost entirely uncorrelated ($r = -0.02$) with the Level 2 residuals.

Part F

Create a scatterplot showing the relationship between the Level 2 residuals and classroom mean extraversion. Does there appear to be any need to model nonlinearity at Level 2?

```
Class_Data %>%
  ggplot(aes(x = L2_residuals, y = Mean_E_GMC)) +
  geom_point() +
  geom_smooth() +
  theme_classic()
```



The relationship between classroom mean extraversion and the Level 2 residuals does appear to be slightly nonlinear.

Question 6

6. Fit a new model based on what you have discovered so far:

To fit this model, you will need to create a new variable, Mean_E_GMC_SQ, that is the square of Mean_E_GMC. Merge both variables into the original data frame, and fit the new model (call it Pop_Fit_3). Use the lme4 package so that you don't encounter convergence problems. Is there any evidence of curvilinearity at Level 2?

```
dat <- Class_Data %>% mutate(Mean_E_GMC_SQ = Mean_E_GMC^2) %>%
  full_join(dat)

Pop_Fit_3 <- lmer(popular ~ extrav*Mean_E_GMC + extrav*Mean_E_GMC_SQ + sex*Mean_E_GMC +
  sex*Mean_E_GMC_SQ + (extrav + sex|class), data = dat)

table_fun(Pop_Fit_3) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
    col.names = c("Term", c("b", "CI"))) %>%
  add_header_above(c(" " = 1, "Fit 3")) %>%
  group_rows("Fixed", 1,9) %>%
  group_rows("Random", 10,12) %>%
  group_rows("Fixed", 13,14)
```

Term	Fit 3	
	b	CI
Fixed		
(Intercept)	20.34	[12.05, 26.86]
extrav	-1.41	[-2.09, -0.32]
Mean E GMC	-26.12	[-38.05, -9.88]
Mean E GMC SQ	7.67	[0.15, 13.56]
sexFemale	0.49	[-2.30, 4.74]
extrav:Mean E GMC	2.66	[0.56, 4.14]
extrav:Mean E GMC SQ	-0.79	[-1.55, 0.20]
Mean E GMC:sexFemale	0.99	[-6.99, 6.15]
Mean E GMC SQ:sexFemale	-0.22	[-2.59, 3.42]
Random		
τ_{00}	1.58	[1.08, 2.19]
τ_{11}	0.02	[0.01, 0.03]
τ_{22}	0.00	[0.00, 0.02]
R^2_m	0.43	
R^2_c	0.70	

There is no evidence of linearity at Level 2 (all t's < 1.6).

Question 7

Fit a model that eliminates the squared terms in Level 2 (call it Pop_Fit_4) and compare it to the full model. Are you justified in eliminating the squared terms?

```
Pop_Fit_4 <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (extrav + sex | class), data = dat)
```

```
table_fun(Pop_Fit_4) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
        col.names = c("Term", c("b", "CI"))) %>%
  add_header_above(c(" " = 1, "Fit 4")) %>%
  group_rows("Fixed", 1,6) %>%
  group_rows("Random", 7,9) %>%
  group_rows("Fixed", 10,11)
```

Term	Fit 4	
	b	CI
Fixed		
(Intercept)	10.84	[8.75, 12.70]
extrav	-0.43	[-0.68, -0.11]
Mean E GMC	-8.82	[-10.75, -6.73]
sexFemale	0.72	[0.22, 1.08]
extrav:Mean E GMC	0.88	[0.56, 1.14]
Mean E GMC:sexFemale	0.53	[0.22, 1.01]
Random		
τ_{00}	1.59	[0.86, 2.19]
τ_{11}	0.02	[0.01, 0.03]
τ_{22}	0.00	[0.00, 0.04]
R^2_m	0.41	
R^2_c	0.69	

```
anova(Pop_Fit_4, Pop_Fit_3)
```

Data: dat Models: Pop_Fit_4: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav + Pop_Fit_4: sex | class) Pop_Fit_3: popular ~ extrav * Mean_E_GMC + extrav * Mean_E_GMC_SQ + sex * Pop_Fit_3: Mean_E_GMC + sex * Mean_E_GMC_SQ + (extrav + sex | class) Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq) Pop_Fit_4 13 4837.8 4910.6 -2405.9 4811.8 Pop_Fit_3 16 4841.2 4930.8 -2404.6 4809.2 2.6108 3 0.4556

Debatably, the model that includes the squared term not better than the one that doesn't. I would drop it.

Question 8

Using the simpler Pop_Fit_4 model, determine if either or both of the slope variances at Level 2 can be set to 0.

```
Pop_Fit_4a <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (extrav | class), data = dat)
Pop_Fit_4b <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (sex | class), data = dat)
Pop_Fit_4c <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (1 | class), data = dat)

f4a.tab <- table_fun(Pop_Fit_4a)
f4b.tab <- table_fun(Pop_Fit_4b)
f4c.tab <- table_fun(Pop_Fit_4c)

f4a.tab %>% mutate(model = "Fit 4a") %>%
  full_join(f4b.tab %>% mutate(model = "Fit 4b")) %>%
  full_join(f4c.tab %>% mutate(model = "Fit 4c")) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  gather(key = est, value = value, b, CI) %>%
  unite(tmp, model, est, sep = ".") %>%
  mutate(type = factor(type, levels = c("Fixed Parts", "Random Parts", "Model Terms"))) %>%
  spread(key = tmp, value = value) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
        col.names = c("Term", rep(c("b", "CI"), times = 3))) %>%
  add_header_above(c(" " = 1, "extrav RE" = 2, "sex RE" = 2, "No RE Slopes" = 2)) %>%
  group_rows("Fixed", 1,6) %>%
```

```
group_rows("Random", 7,8) %>%
group_rows("Fixed", 9,10)
```

Term	extrav RE		sex RE		No RE Slopes	
	b	CI	b	CI	b	CI
Fixed						
(Intercept)	10.86	[8.67, 13.60]	10.96	[7.51, 11.73]	10.97	[9.51, 12.28]
extrav	-0.43	[-0.57, -0.22]	-0.43	[-0.58, -0.14]	-0.43	[-0.57, -0.25]
extrav:Mean E GMC	0.88	[0.69, 1.02]	0.88	[0.58, 1.05]	0.88	[0.69, 1.01]
Mean E GMC	-8.85	[-11.55, -6.62]	-8.94	[-9.72, -5.41]	-8.96	[-10.07, -7.51]
Mean E GMC:sexFemale	0.55	[0.04, 1.12]	0.47	[-0.36, 0.99]	0.47	[0.05, 1.16]
sexFemale	0.70	[0.18, 1.18]	0.78	[0.22, 1.57]	0.78	[0.06, 1.25]
Random						
τ_{00}	1.60	[1.23, 2.02]	0.49	[0.33, 0.62]	0.46	[0.39, 0.59]
τ_{11}	0.02	[0.02, 0.03]	0.00	[0.00, 0.02]		
R^2_c	0.69		0.68		0.68	
R^2_m	0.41		0.41		0.41	

```
anova(Pop_Fit_4, Pop_Fit_4a, Pop_Fit_4b, Pop_Fit_4c)
```

Data: dat Models: Pop_Fit_4c: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (1 | class)
Pop_Fit_4a: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav | Pop_Fit_4a: class)
Pop_Fit_4b: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (sex | class) Pop_Fit_4:
popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav + Pop_Fit_4: sex | class) Df AIC
BIC logLik deviance Chisq Chi Df Pr(>Chisq)
Pop_Fit_4c 8 4865.8 4910.6 -2424.9 4849.8
Pop_Fit_4a 10 4833.0 4889.0 -2406.5 4813.0 36.802 2 1.020e-08 **Pop_Fit_4b 10 4869.1 4925.1 -2424.6**
4849.1 0.000 0 1
Pop_Fit_4 13 4837.8 4910.6 -2405.9 4811.8 37.315 3 3.947e-08 — Signif. codes: 0 ‘’ 0.001 ‘’
0.01 ‘’ 0.05 ‘’ 0.1 ‘’ 1

Part A

Eliminate the random effect for extrav (call this model Pop_Fit_5). Is this model indistinguishable from Pop_Fit_4?

```
Pop_Fit_5 <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (sex | class), data = dat)
```

```
table_fun(Pop_Fit_5) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
        col.names = c("Term", c("b", "CI"))) %>%
  add_header_above(c(" " = 1, "Fit 5")) %>%
  group_rows("Fixed", 1,6) %>%
  group_rows("Random", 7,8) %>%
  group_rows("Fixed", 9,10)
```


Term	Fit 5	
	b	CI
Fixed		
(Intercept)	10.96	[9.75, 11.81]
extrav	-0.43	[-0.60, -0.20]
Mean E GMC	-8.94	[-9.62, -7.52]
sexFemale	0.78	[0.15, 1.53]
extrav:Mean E GMC	0.88	[0.62, 1.05]
Mean E GMC:sexFemale	0.47	[-0.23, 1.15]
Random		
τ_{00}	0.49	[0.34, 0.66]
τ_{11}	0.00	[0.00, 0.05]
R^2_m	0.41	
R^2_c	0.68	

```
anova(Pop_Fit_4, Pop_Fit_5)
```

Data: dat Models: Pop_Fit_5: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (sex | class)
 Pop_Fit_4: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav + Pop_Fit_4: sex | class)
 Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
 Pop_Fit_5 10 4869.1 4925.1 -2424.6 4849.1
 Pop_Fit_4 13 4837.8 4910.6 -2405.9 4811.8 37.315 3 3.947e-08 *** — Signif. codes: 0 ‘’ **0.001** ’’ 0.01 ’’ 0.05 ‘’ 0.1 ‘’ 1

This model is distinguishable from the model from question 4 – eliminating the extraversion random slope improves model fit.

Part B

Eliminate the random effect for sex (call this model Pop_Fit_6). Is this model indistinguishable from Pop_Fit_4?

```
Pop_Fit_6 <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (extrav | class), data = dat)

table_fun(Pop_Fit_6) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
        col.names = c("Term", c("b", "CI"))) %>%
  add_header_above(c(" " = 1, "Fit 6")) %>%
  group_rows("Fixed", 1,6) %>%
  group_rows("Random", 7,8) %>%
  group_rows("Fixed", 9,10)
```

Term	Fit 6	
	b	CI
Fixed		
(Intercept)	10.86	[8.71, 12.03]
extrav	-0.43	[-0.56, -0.23]
Mean E GMC	-8.85	[-10.05, -6.60]
sexFemale	0.70	[0.02, 1.53]
extrav:Mean E GMC	0.88	[0.68, 1.00]
Mean E GMC:sexFemale	0.55	[-0.29, 1.21]
Random		
τ_{00}	1.60	[1.29, 1.94]
τ_{11}	0.02	[0.01, 0.03]
R^2_m	0.41	
R^2_c	0.69	

```
anova(Pop_Fit_4, Pop_Fit_6)
```

Data: dat Models: Pop_Fit_6: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav | Pop_Fit_6: class) Pop_Fit_4: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav + Pop_Fit_4: sex | class) Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq) Pop_Fit_6 10 4833.0 4889.0 -2406.5 4813.0 Pop_Fit_4 13 4837.8 4910.6 -2405.9 4811.8 1.247 3 0.7418

Yes, the model that does not include the random slope for sex is indistinguishable from one that does.

Part C

Finally, eliminate them both (call this Pop_Fit_7) and compare it to Pop_Fit_4. Which simpler model is justified?

```
Pop_Fit_7 <- lmer(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC + (1 | class), data = dat)

table_fun(Pop_Fit_7) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
        col.names = c("Term", c("b", "CI"))) %>%
  add_header_above(c(" " = 1, "Fit 7")) %>%
  group_rows("Fixed", 1,6) %>%
  group_rows("Random", 7,7) %>%
  group_rows("Fixed", 8,9)
```

Term	Fit 7	
	b	CI
Fixed		
(Intercept)	10.97	[8.75, 12.01]
extrav	-0.43	[-0.66, -0.23]
Mean E GMC	-8.96	[-10.02, -6.71]
sexFemale	0.78	[0.24, 1.48]
extrav:Mean E GMC	0.88	[0.69, 1.11]
Mean E GMC:sexFemale	0.47	[-0.23, 1.01]
Random		
τ_{00}	0.46	[0.43, 0.60]
R^2_m	0.41	
R^2_c	0.68	

```
anova(Pop_Fit_4, Pop_Fit_7)
```

```
## Data: dat
## Models:
## Pop_Fit_7: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (1 | class)
## Pop_Fit_4: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (extrav +
## Pop_Fit_4:      sex | class)
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## Pop_Fit_7  8 4865.8 4910.6 -2424.9   4849.8
## Pop_Fit_4 13 4837.8 4910.6 -2405.9   4811.8 38.049      5 3.689e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(Pop_Fit_5, Pop_Fit_7)
```

```
## Data: dat
## Models:
## Pop_Fit_7: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (1 | class)
## Pop_Fit_5: popular ~ extrav * Mean_E_GMC + sex * Mean_E_GMC + (sex | class)
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## Pop_Fit_7  8 4865.8 4910.6 -2424.9   4849.8
## Pop_Fit_5 10 4869.1 4925.1 -2424.6   4849.1 0.7339      2    0.6928
```

The simplest model with no random slopes is best because it is better than a model that includes both but not better than one that includes only sex.

Question 9

Use the simplest model justified from the previous question.

Part A

Retest the Level 1 homogeneity assumption. Is there any improvement compared to what was found for Question 1?

```
Pop_Fit_7.lme <- lme(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC, random = ~ 1|class, data = dat)
Pop_Fit_7.a <- lme(popular ~ extrav*Mean_E_GMC + sex*Mean_E_GMC, random = ~ 1|class,
  data = dat, varIdent(form = ~ 1 | class))
```

```
anova(Pop_Fit_7.lme, Pop_Fit_7.a)
```

```
##           Model  df      AIC      BIC    logLik   Test  L.Ratio
## Pop_Fit_7.lme    1    8 4882.758 4927.541 -2433.379
## Pop_Fit_7.a      2 107 4963.523 5562.499 -2374.762 1 vs 2 117.2343
##           p-value
## Pop_Fit_7.lme
## Pop_Fit_7.a      0.1019
```

In this case, we meet the homogeneity of variance assumption. The model that fits unique variances is not better than one that does not.

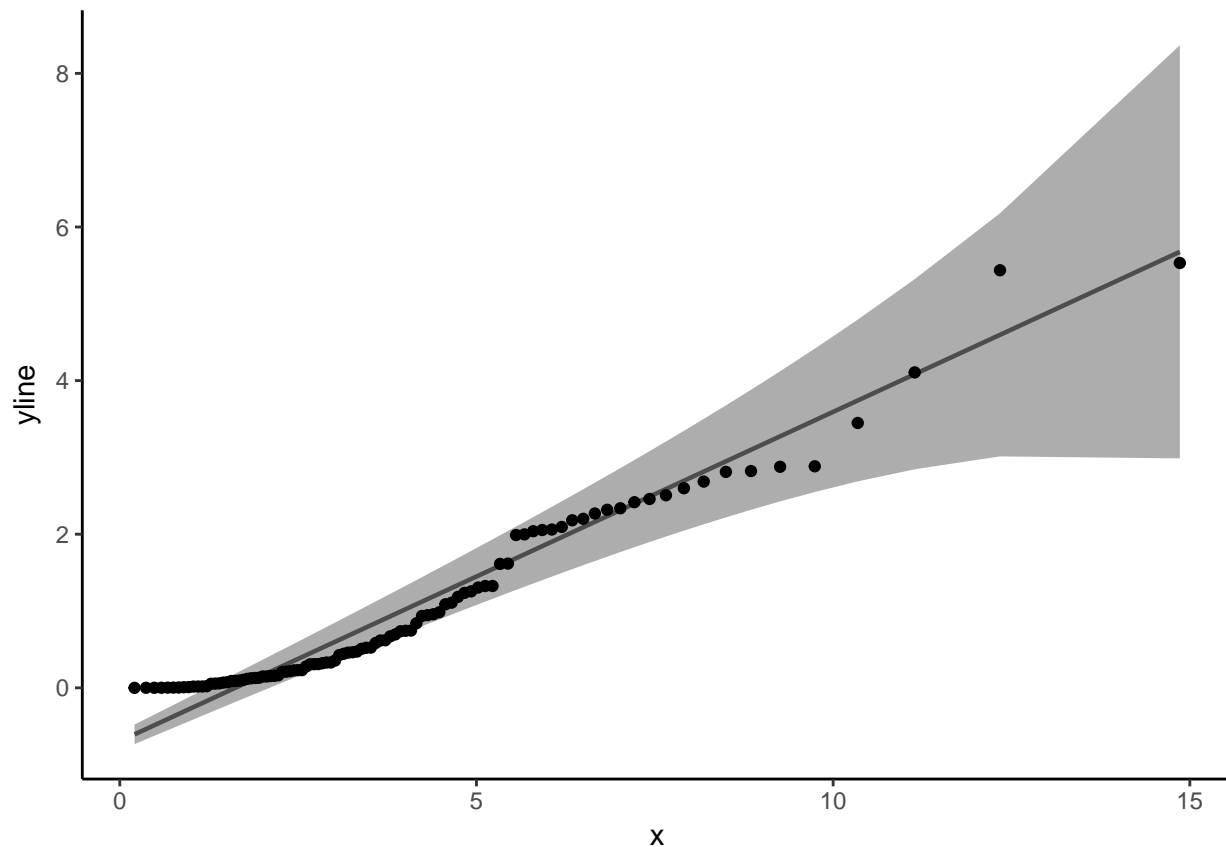
Part B

Check the multivariate normality assumption for the residuals at Level 2 using Mahalanobis distance. Are the residuals multivariate normal?

```
L2_residuals <- ranef(Pop_Fit_7)[[1]]
MD.v <- mahalanobis(L2_residuals, colMeans(L2_residuals), cov(L2_residuals))

MD.df <- MD.v %>% data.frame %>% setNames("MD") %>% mutate(class = names(MD.v))

MD.df %>%
  ggplot(aes(sample = MD)) +
  stat_qq_band(distribution = "chisq", dparams = list(df = 4)) +
  stat_qq_line(distribution = "chisq", dparams = list(df = 4)) +
  stat_qq_point(distribution = "chisq", dparams = list(df = 4)) +
  theme_classic()
```



The residuals do not appear to be multivariate normal.

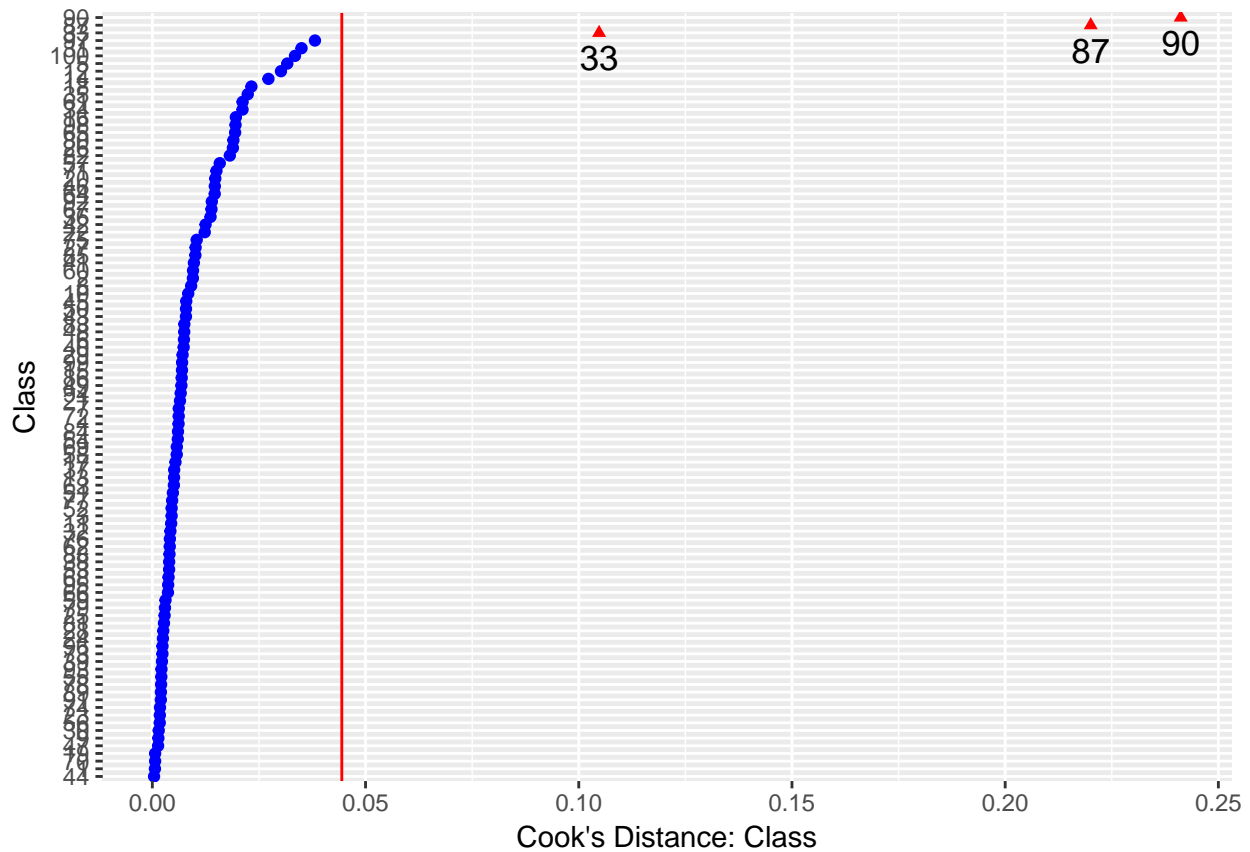
Question 10

Use Cook's distance to examine how influential the classrooms are in the model fit for Question 9.

```
Pop_Fit_7.i <- influence(Pop_Fit_7, group = "class")

Cooks_Class <- cooks.distance(Pop_Fit_7.i, group = "class")
Cooks_Class <- Cooks_Class %>% data.frame %>% setNames("cooks.distance") %>%
  mutate(class = rownames(Cooks_Class))

dotplot_diag(x = cooks.distance, index = class, data = Cooks_Class,
  cutoff = "internal", name = "cooks.distance",
  ylab = "Cook's Distance: Class", xlab = "Class")
```



Part A

How many classrooms stand out as distinctly more influential than the others? 3 classrooms: 33, 87, and 90.

Part B

If those classrooms are excluded from the analysis and the model is refit, do any conclusions change?

```
Pop_Fit_7.cd <- update(Pop_Fit_7, data = dat %>% filter(!(class %in% c(87, 33, 90))))

table_fun(Pop_Fit_7.cd) %>%
  mutate(term = str_replace_all(term, "_", " ")) %>%
  select(-type) %>%
  kable(., "latex", escape = F, booktabs = T,
        col.names = c("Term", c("b", "CI"))) %>%
  add_header_above(c(" " = 1, "Fit 7 Outliers Removed")) %>%
  group_rows("Fixed", 1,6) %>%
  group_rows("Random", 7,7) %>%
  group_rows("Fixed", 8,9)
```

Term	Fit 7 Outliers Removed	
	b	CI
Fixed		
(Intercept)	13.17	[12.00, 14.94]
extrav	-0.62	[-0.80, -0.27]
Mean E GMC	-11.26	[-13.08, -10.04]
sexFemale	0.53	[-0.24, 0.97]
extrav:Mean E GMC	1.08	[0.73, 1.24]
Mean E GMC:sexFemale	0.73	[0.28, 1.50]
Random		
τ_{00}	0.44	[0.36, 0.54]
R^2_m	0.43	
R^2_c	0.68	

No, in terms of significance levels, no conclusions change.