Final

```
knitr::opts_chunk$set(echo = TRUE)
library(faraway)
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
```

```
library(tidyr)
library(lme4)
```

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

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lattice)
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:faraway':
##
##     melanoma
```

```
library(RLRsim)
library(pbkrtest)
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(geepack)
```

```
##
## Attaching package: 'geepack'
```

```
## The following object is masked from 'package:faraway':
##
##     ohio
```

```
library(INLA)
```

```
## Loading required package: sp
```

```
## Loading required package: parallel
```

```
## Loading required package: foreach
```

```
## This is INLA_20.03.17 built 2020-10-30 17:30:50 UTC.
## See www.r-inla.org/contact-us for how to get help.
## To enable PARDISO sparse library; see inla.pardiso()
```

##Question 1 1. A “round robin” study is one where the same experiment is performed by a number of different labs, in order to assess how well the different labs are able to reproduce each others’ work. As part of such a study, seven labs are asked to conduct tensile strength measurements on samples of steel wire. In total, 44 such measurements are made. The file RoundRobin.csv contains the raw data, which are summarize in Table 1. *similar to pulp example from chp. 10*

```
rr=read.csv('/Users/SylviaSzarka/Desktop/School/STOR 590/FINAL/RoundRobin.csv')
```

a. Fit a simple linear regression model with Strength as the response and Lab as a predictor. Is the Lab effect statistically significant? [3 points]

```
op <- options(contrasts=c("contr.sum","contr.poly"))
options(op)
#Treats Lab as factor variable
slr<-lm(Strength~Lab,rr)
summary(slr)
```

```
##
## Call:
## lm(formula = Strength ~ Lab, data = rr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.960  -4.367  -0.502   4.283  39.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   85.460      5.023  17.013   <2e-16 ***
## LabB          -3.280      7.104  -0.462   0.6470
## LabC           3.365      6.403   0.526   0.6024
## LabD          -5.403      6.577  -0.822   0.4166
## LabE           4.040      7.535   0.536   0.5950
## LabF          12.807      6.265   2.044   0.0481 *
## LabG          -2.593      6.801  -0.381   0.7052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.23 on 37 degrees of freedom
## Multiple R-squared:  0.2729, Adjusted R-squared:  0.155
## F-statistic: 2.315 on 6 and 37 DF,  p-value: 0.05357
```

```
#Analysis of variance table
amod<-aov(Strength~Lab,rr)
summary(amod)
```

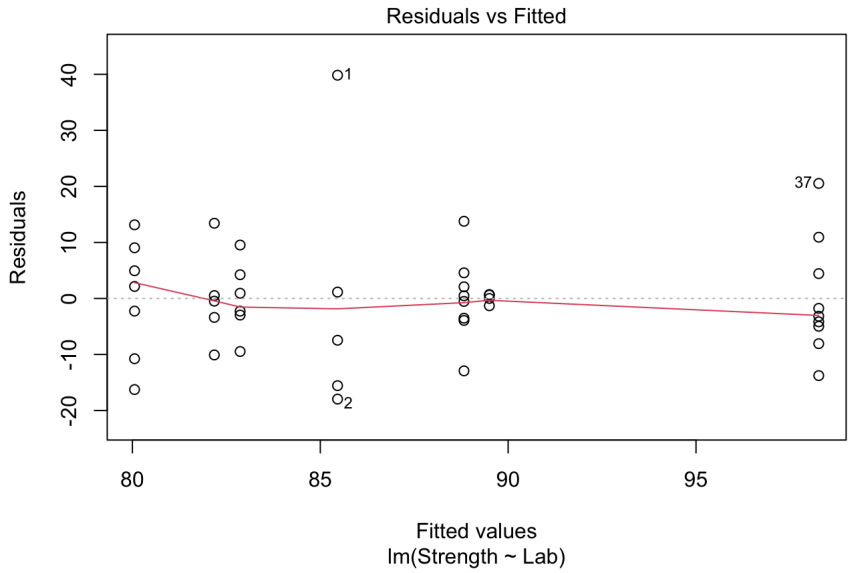
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Lab          6   1752   292.0   2.315 0.0536 .
## Residuals   37   4668   126.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

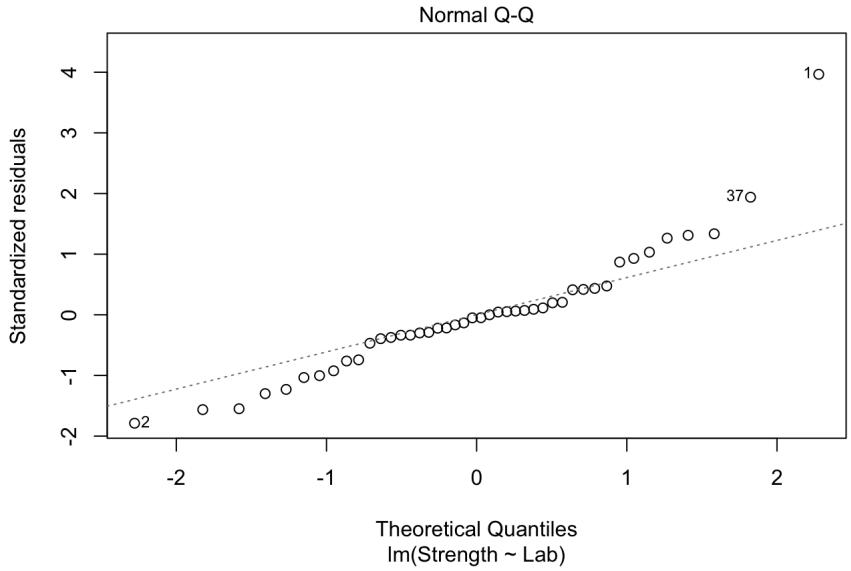
```
#P-value of 0.0536-> we fail to reject the Null hypothesis that there is no regression effect, meaning that there is no lab effect
#1752: Sum sq due to regression
#1.70: Sum sq. due to errors
```

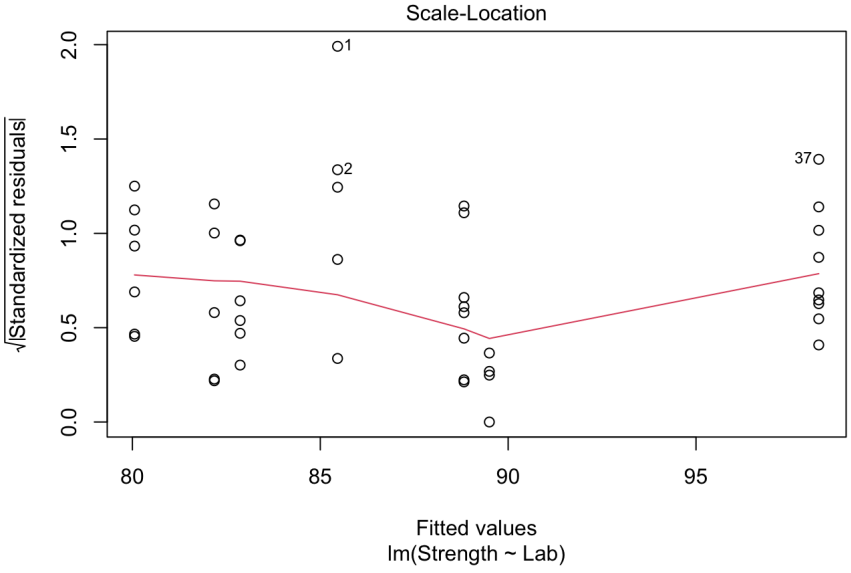
Out of the 7 labs, only 2 of them are statistically significant - Lab A (the intercept in the model), and lab F. Most of the Standard errors are greater than the coefficient estimates, other than the 2 significant Lab effects. The p-value of 0.0536 for the combined lab effects suggests that the lab effect is not significant overall (and there is no significant difference in the measurements between the Labs).

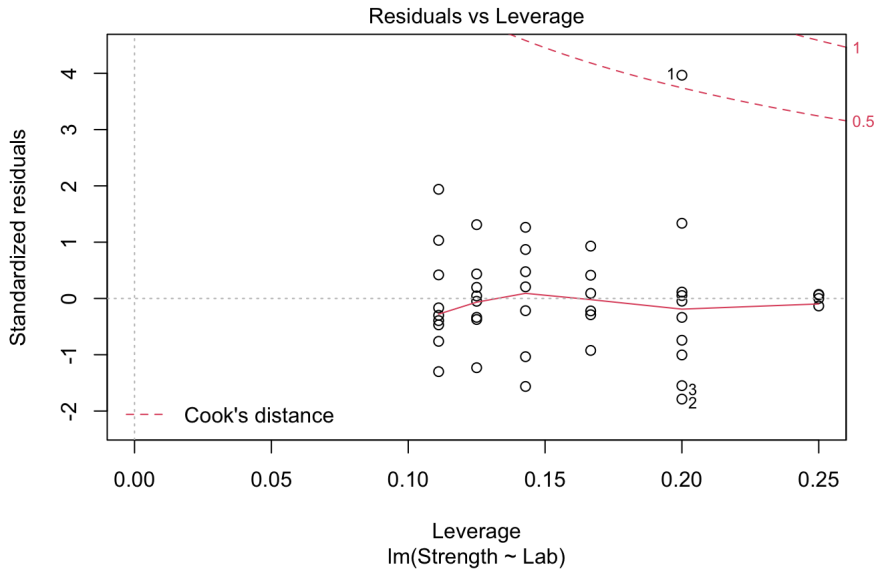
b. Are there any outliers in the data? Use standard diagnostics to determine which (if any) observations might be outliers, giving your reasons. Rerun the analysis without the suspected outliers and state any changes from your conclusions in (a). [3 points]

```
plot(slr)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Strength ~ Lab)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Strength ~ Lab)

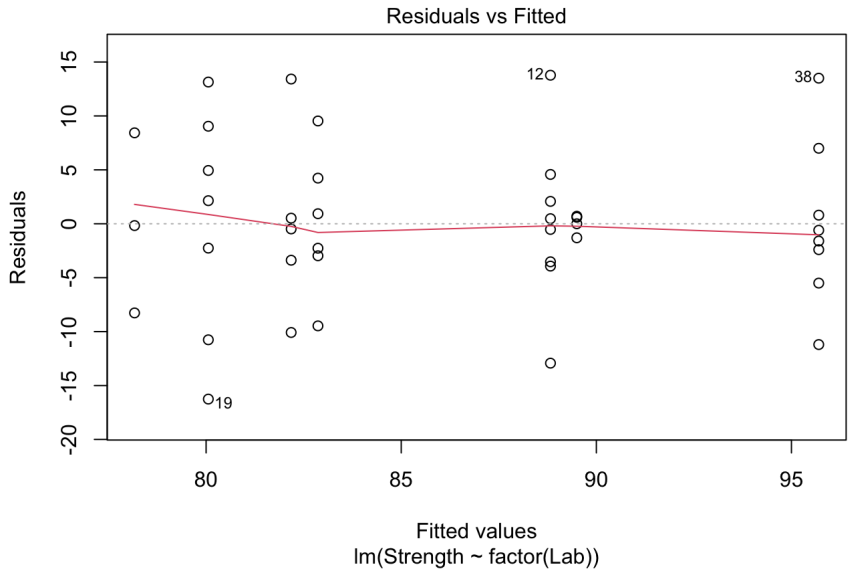

Scale-Location
√|Standardized residuals|
Fitted values
lm(Strength ~ Lab)

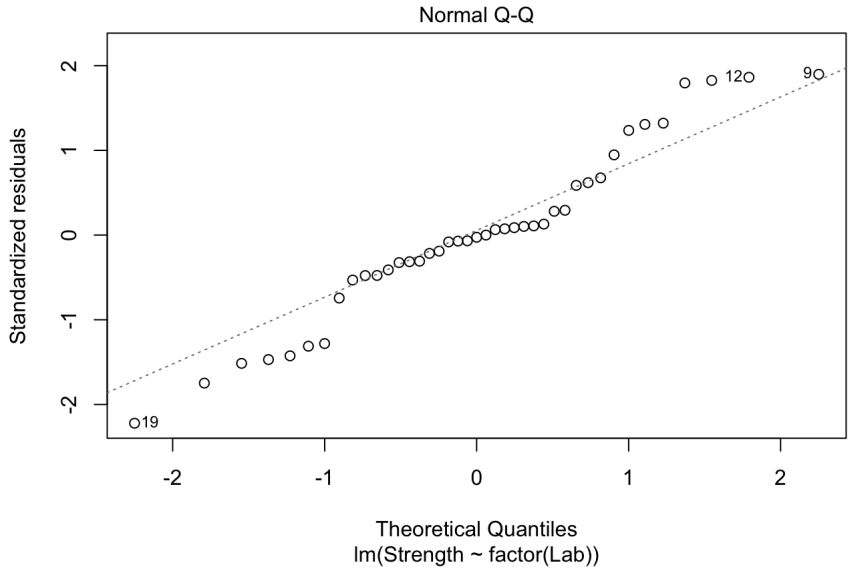


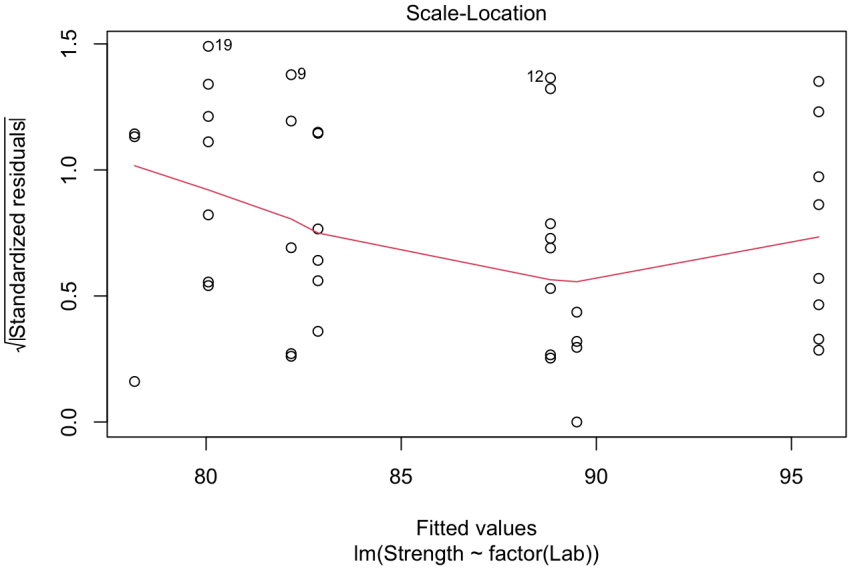

```
#Omitting all potential outliers:
slromit<-lm(Strength~factor(Lab),rr,subset=c(-1,-2,-37))
summary(slromit)
```

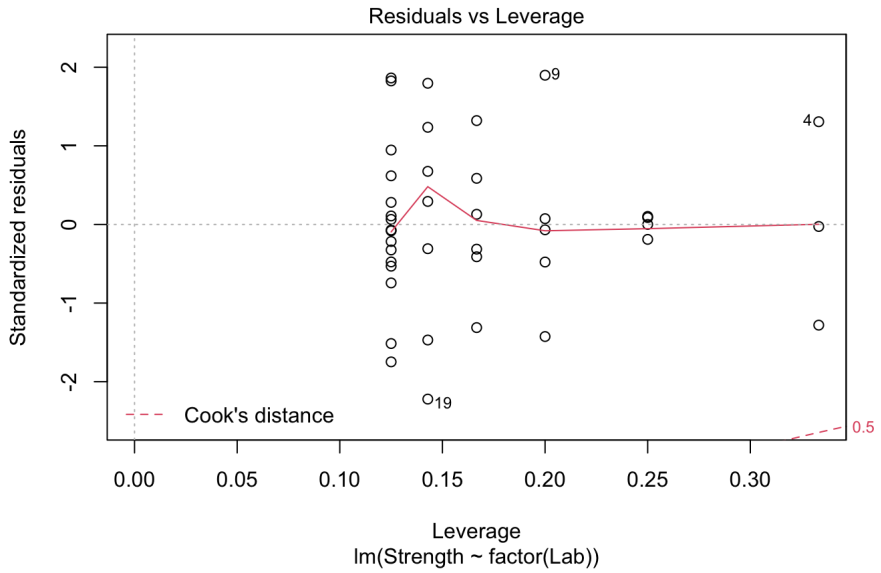
```
##
## Call:
## lm(formula = Strength ~ factor(Lab), data = rr, subset = c(-1,
##     -2, -37))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2571  -3.3800  -0.1667   4.2333  13.7750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    78.167      4.563  17.129  < 2e-16 ***
## factor(Lab)B    4.013      5.772   0.695  0.49161
## factor(Lab)C   10.658      5.351   1.992  0.05447 .
## factor(Lab)D    1.890      5.454   0.347  0.73103
## factor(Lab)E   11.333      6.037   1.877  0.06907 .
## factor(Lab)F   17.533      5.351   3.277  0.00242 **
## factor(Lab)G    4.700      5.589   0.841  0.40626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.904 on 34 degrees of freedom
## Multiple R-squared:  0.4016, Adjusted R-squared:  0.296
## F-statistic: 3.803 on 6 and 34 DF,  p-value: 0.005255
```

```
#Combined p-value changes dramatically from not significant (0.0536) to highly significant (0.005255); Interestin
gly Lab D has an even less significant p-value while the others increase more as compared to before.
plot(slromit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Strength ~ factor(Lab))
12
38
19


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Strength ~ factor(Lab))
12
9
19

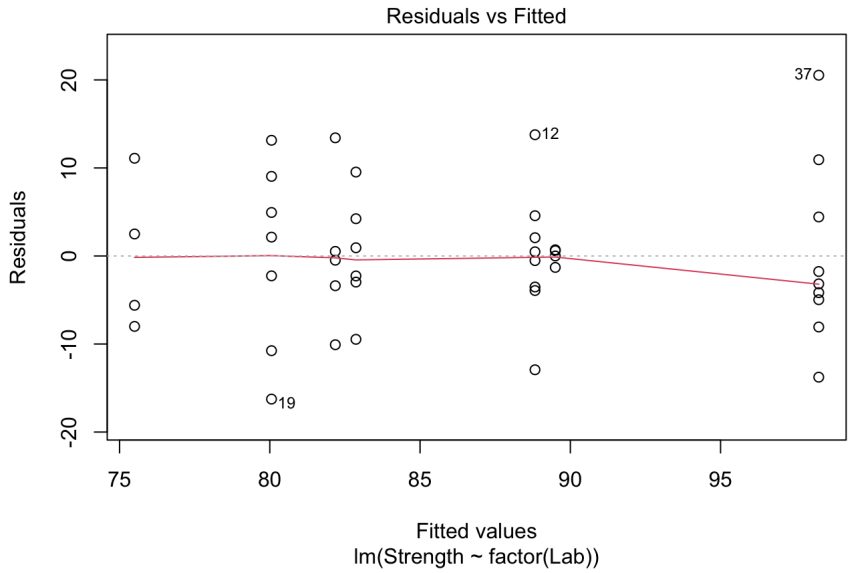

Scale-Location
√|Standardized residuals|
Fitted values
lm(Strength ~ factor(Lab))
19
9
12

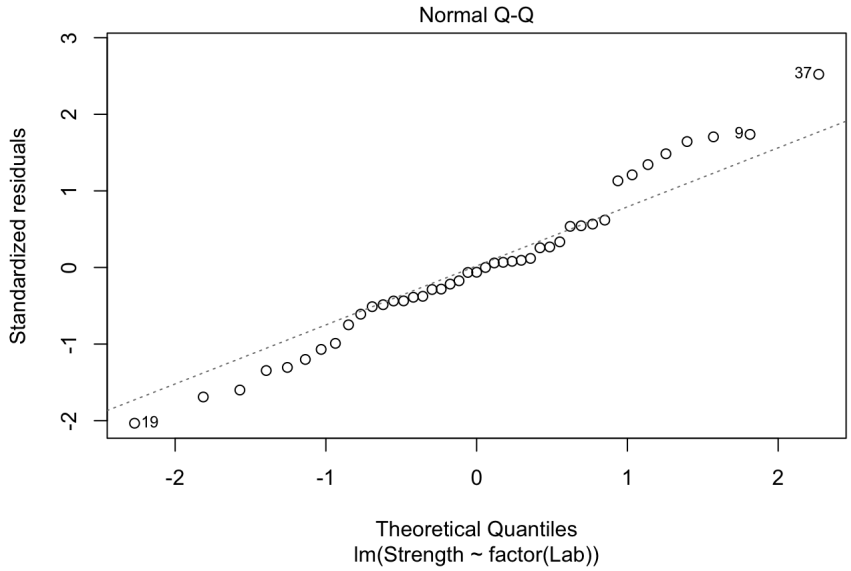


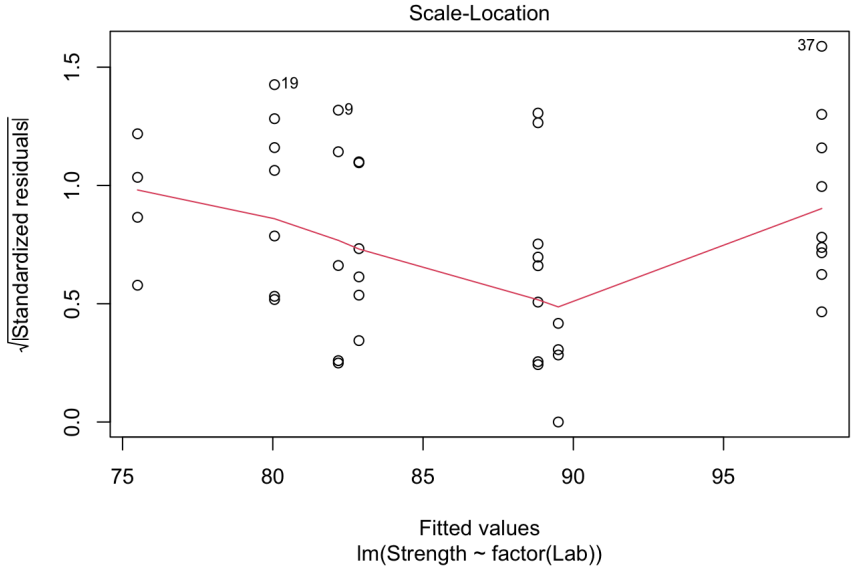

```
#However, after plotting the dataset after omitting the potential 3 outliers, we now have 3 new outliers.
#Omitting just 1 (most influential) outlier:
slromit1<-lm(Strength~factor(Lab),rr,subset=c(-1))
summary(slromit1)
```

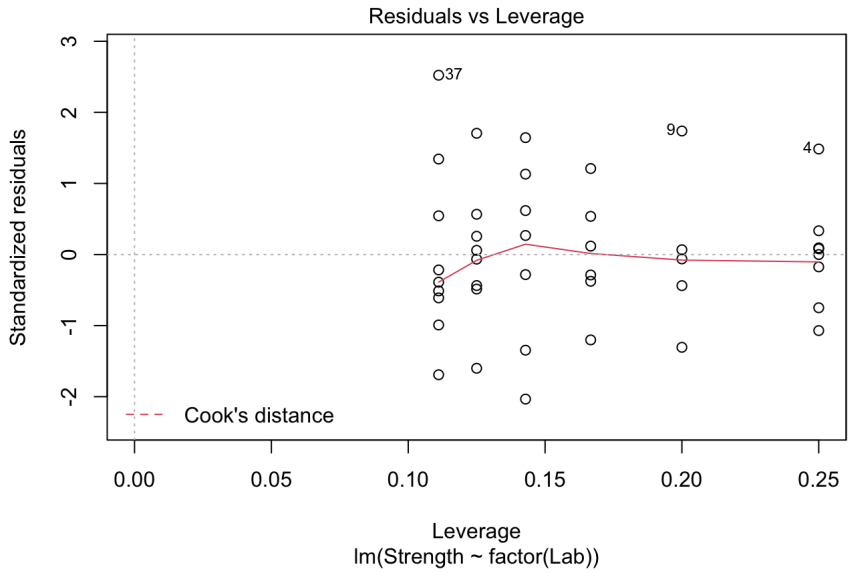
```
##
## Call:
## lm(formula = Strength ~ factor(Lab), data = rr, subset = c(-1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.257  -4.046  -0.480   4.333  20.533
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    75.500      4.317  17.488  < 2e-16 ***
## factor(Lab)B    6.680      5.792   1.153   0.2564
## factor(Lab)C   13.325      5.287   2.520   0.0163 *
## factor(Lab)D    4.557      5.412   0.842   0.4053
## factor(Lab)E   14.000      6.105   2.293   0.0278 *
## factor(Lab)F   22.767      5.189   4.388 9.58e-05 ***
## factor(Lab)G    7.367      5.573   1.322   0.1946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.634 on 36 degrees of freedom
## Multiple R-squared:  0.4585, Adjusted R-squared:  0.3682
## F-statistic: 5.079 on 6 and 36 DF,  p-value: 0.0007252
```

```
#Overall lab effect becomes even more statistically significant with a p-value = 0.00072 and now Lab C and E are
 also significant now on top of Labs A & F.
plot(slromit1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Strength ~ factor(Lab))
37
12
19


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Strength ~ factor(Lab))
37
9
19


Scale-Location
√|Standardized residuals|
Fitted values
lm(Strength ~ factor(Lab))
37
19
9




```
#The Normal Q-Q plot looks  more normally distributed and this model looks to fit the data better than previously
when omitting all 3 potential outliers. However, there are still 3 more outliers that  appear again after remova
l, and this normal Q-Q plot compared to the original slr model is not improved by that much.
```

Points 1, 2, and 37 appear to be potential outliers on the Residual v. Fitted & normal Q-Q plot (with Point 1 appearing most influential). However, after removing the potential outliers, there were more outliers that seemed to appear. This could be due to the small sample size, and in this case it seems that removing the outliers that will just create more outliers so it is not satisfactory to keep removing them to create more. The removal of one influential point caused a dramatic increase in the p-value, making the lab effect highly statistically significant. This dramatic change seemed problematic because it made some labs statistically significant when they were not before. Taking out this outlier that corresponds to Lab A's measurement would be removing a measurement that is far off from the other Labs, which would be benefit Lab A, and potentially make other deviations in measurements from other labs look larger. That is probably why the lab effect increased as well. So, unless we are convinced that the datapoint is truly an error, it is not wise to remove any points, especially when we are trying to compare differences in the performance of the Labs with a small sample size.

c. Now run this as a random effects regression using lme4, without removing the outlier. State the estimated standard deviations for both the Lab effect and the residual, and calculate a 95% confidence interval for each. [6 points]

```
rr$Lab<-as.factor(rr$Lab)
rer<-lmer(Strength~1+(1|Lab),rr)
summary(rer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Strength ~ 1 + (1 | Lab)
##    Data: rr
##
## REML criterion at convergence: 338
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7141 -0.4385 -0.0060  0.2226  3.5071
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Lab      (Intercept)  25.51    5.051
##  Residual             124.19   11.144
## Number of obs: 44, groups:  Lab, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   86.993      2.565   33.91
```

```
#95% confidence interval
confint(rer)
```

```
## Computing profile confidence intervals ...
```

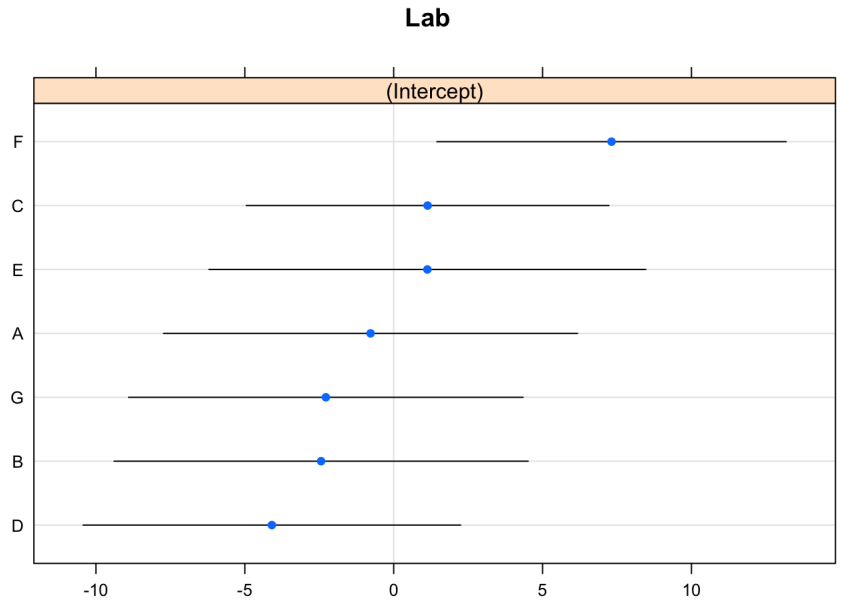
```
##                 2.5 %   97.5 %
## .sig01       0.000000 10.63936
## .sigma       9.025679 14.15326
## (Intercept) 81.566579 92.24825
```

The intercept is 86.99. The standard deviation for the Lab effect is 5.051, while the standard deviation for the residual is 11.144. The 95% confidence interval for the Lab effect (sig01) is 0 to 10.64. The 95% confidence interval for the residual effect (sigma squared epsilon) is 9.03 to 14.15 [effect due to error].

d. Draw a lattice plot to show the means and confidence intervals for the seven lab effects. [4 points]

```
dotplot(ranef(rer,condVar=TRUE))
```

```
## $Lab
```



e. Now run this as a Bayesian analysis using either STAN or INLA (your choice!). Fit a suitable model for a one-way analysis of variance, and show the following:

i. A plot of posterior densities for the Lab and Residual standard deviations; [3 points]
ii. A plot of posterior densities for the seven Lab effects; [3 points]
iii. A summary table of posterior distributions for the main parameters of the models. [4 points] *CHP 12*

```
#USING INLA.
formula=Strength~f(Lab,model="iid")
result=inla(formula,family="gaussian",data=rr)
summary(result)
```

```
##
## Call:
##    "inla(formula = formula, family = \"gaussian\", data = rr)"
## Time used:
##     Pre = 2.72, Running = 0.25, Post = 0.135, Total = 3.11
## Fixed effects:
##               mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) 87.473 1.849     83.825   87.473     91.116 87.473   0
##
## Random effects:
##   Name     Model
##     Lab IID model
##
## Model hyperparameters:
##                                           mean       sd 0.025quant 0.5quant
## Precision for the Gaussian observations 7.00e-03 2.00e-03      0.004 7.00e-03
## Precision for Lab                       1.86e+04 1.81e+04   1265.699 1.33e+04
##                                         0.975quant     mode
## Precision for the Gaussian observations       0.01    0.007
## Precision for Lab                         66875.95 3455.616
##
## Expected number of effective parameters(stdev): 1.00(0.001)
## Number of equivalent replicates : 43.99
##
## Marginal log-Likelihood:  -186.22
```

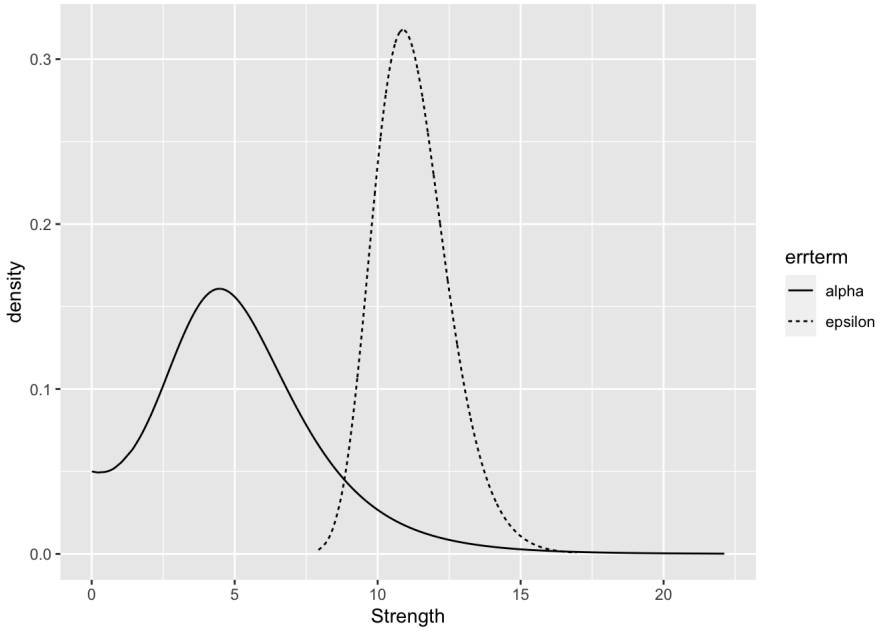
```
#Intercept: 87.04
#Precision for gaussian observation (sigma epsilon): 0.007
#Posterior mean for sigma epsilon: (1/sqrt(0.007)) = 11.95
#Precision for Lab (sigma alpha): 18600, posterior mean for sigma alpha is close to 0.. try hyperprior

#Adding hyperprior
sdres <- sd(rr$Strength)
pcprior <- list(prec = list(prior="pc.prec", param = c(3*sdres,0.01)))
formula <- Strength ~ f(Lab, model="iid", hyper = pcprior)
result <- inla(formula, family="gaussian", data=rr)
result <- inla.hyperpar(result)
summary(result)
```

```
##
## Call:
##    "inla(formula = formula, family = \"gaussian\", data = rr)"
## Time used:
##     Pre = 1.43, Running = 0.344, Post = 0.104, Total = 1.88
## Fixed effects:
##              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) 87.04 2.828     81.273   87.078     92.595 87.146   0
##
## Random effects:
##   Name     Model
##     Lab IID model
##
## Model hyperparameters:
##                                           mean       sd 0.025quant 0.5quant
## Precision for the Gaussian observations    0.008 2.00e-03      0.005    0.008
## Precision for Lab                       6086.297 5.93e+06      0.007    0.042
##                                         0.975quant  mode
## Precision for the Gaussian observations      0.012 0.008
## Precision for Lab                            3.942 0.017
##
## Expected number of effective parameters(stdev): 4.03(1.45)
## Number of equivalent replicates : 10.93
##
## Marginal log-Likelihood:  -185.56
```

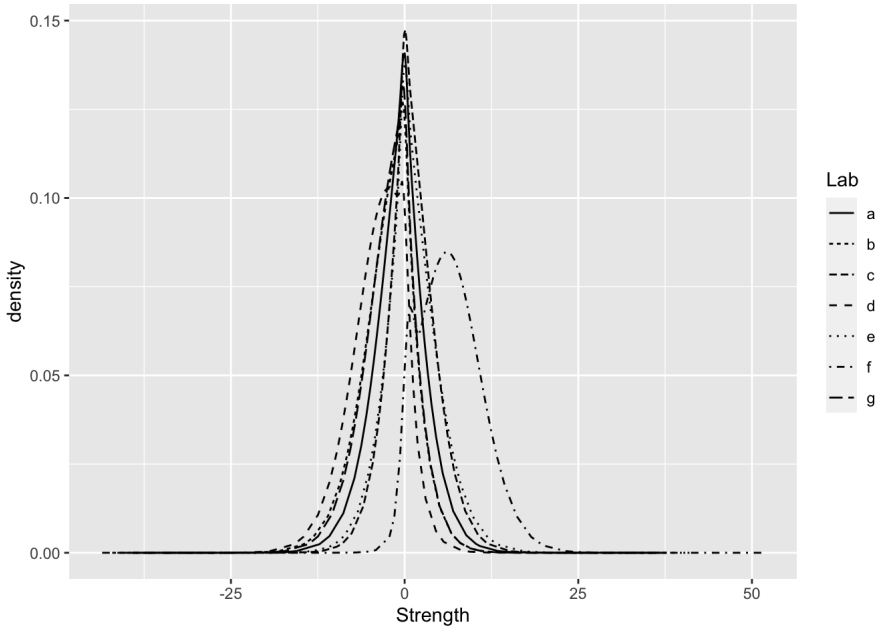
```
#Precision for Gaussian Obs: 0.008
#precision for lab: 6086.297
#posterior mean: 0.0128, not as close to 0 now

#(i). A plot of posterior densities for the Lab and Residual standard deviations;
sigmaalpha <- inla.tmarginal(function(x) 1/sqrt(exp(x)),result$internal.marginals.hyperpar[[2]])
sigmaepsilon <- inla.tmarginal(function(x) 1/sqrt(exp(x)),result$internal.marginals.hyperpar[[1]])
ddf <- data.frame(rbind(sigmaalpha,sigmaepsilon),errterm=gl(2,2048,labels = c("alpha","epsilon")))
ggplot(ddf, aes(x,y, linetype=errterm))+geom_line()+xlab("Strength")+ylab("density")
```




```
#alpha epsilon still somewhat stacked up at 0.

#(ii). A plot of posterior densities for the 7 Lab effects
rdf <- do.call(rbind.data.frame, result$marginals.random$Lab)
rdf <- cbind(Lab=gl(7,nrow(rdf)/7,labels=letters[1:7]),rdf)
ggplot(rdf, aes(x=x,y=y,linetype=Lab))+geom_line()+xlab("Strength")+ylab("density")
```




```
#considerable overlap between densities; hard to distinguish between specific Labs.
#Lab F  seems to vary more than the rest (also has the largest s.d. & mean)

#(iii). A summary table of posterior distributions for the main parameters of the models.
restab <- sapply(result$marginals.fixed, function(x) inla.zmarginal(x,silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaalpha,silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaepsilon,silent=TRUE))
restab <- cbind(restab, sapply(result$marginals.random$Lab,function(x) inla.zmarginal(x, silent=TRUE)))
colnames(restab) = c("mu","alpha","epsilon",levels(rr$Lab))
data.frame(restab)
```

```
##                  mu     alpha  epsilon          A         B         C
## mean       87.04017  5.185612 11.23106 -0.7079821 -2.263979  1.049369
## sd         2.828123  2.867854 1.301421   3.850351  3.982021  3.591792
## quant0.025 81.26006 0.5024386 9.011865  -8.831088 -11.02352 -5.940683
## quant0.25  85.31015  3.230955 10.30365  -2.949431   -4.6248 -1.104513
## quant0.5   87.06452  4.859158 11.11298 -0.5539247  -1.87385 0.7907357
## quant0.75  88.78243  6.727831 12.02753   1.510982 0.2406533  3.123454
## quant0.975 92.57539  11.89789 14.11094    6.96121  5.000824  8.693072
##                     D         E          F         G
## mean        -3.712994  1.096536   6.513965 -2.081521
## sd           3.972397  4.027179   4.497111  3.833641
## quant0.025  -12.46978 -6.719372 -0.6429648  -10.4484
## quant0.25   -6.170822 -1.288837   3.076998 -4.362797
## quant0.5    -3.324217 0.7726311   6.215289 -1.734693
## quant0.75  -0.8725553  3.373704   9.433132  0.301471
## quant0.975   2.995136  9.773656   16.08724  5.006673
```

```
restab
```

```
##            mu       alpha     epsilon  A          B         C
## mean       87.04017 5.185612  11.23106 -0.7079821 -2.263979 1.049369
## sd         2.828123 2.867854  1.301421 3.850351   3.982021  3.591792
## quant0.025 81.26006 0.5024386 9.011865 -8.831088  -11.02352 -5.940683
## quant0.25  85.31015 3.230955  10.30365 -2.949431  -4.6248   -1.104513
## quant0.5   87.06452 4.859158  11.11298 -0.5539247 -1.87385  0.7907357
## quant0.75  88.78243 6.727831  12.02753 1.510982   0.2406533 3.123454
## quant0.975 92.57539 11.89789  14.11094 6.96121    5.000824  8.693072
##            D          E         F          G
## mean       -3.712994  1.096536  6.513965   -2.081521
## sd         3.972397   4.027179  4.497111   3.833641
## quant0.025 -12.46978  -6.719372 -0.6429648 -10.4484
## quant0.25  -6.170822  -1.288837 3.076998   -4.362797
## quant0.5   -3.324217  0.7726311 6.215289   -1.734693
## quant0.75  -0.8725553 3.373704  9.433132   0.301471
## quant0.975 2.995136   9.773656  16.08724   5.006673
```

```
#The standard deviations of the Lab effects are greater than their means..except for Lab F which has the largest
 mean and sd (compared to the other labs)
```

In the density plot in (i), the density for sigma alpha is concentrated near 0. There is a considerable overall in the posterior densities for the 7 lab effects, although Lab F seems to vary more than the others with a higher mean and standard deviation. This is also exemplified in the summary table, which shows that means for the lab effects are smaller than their standard deviations. All of the lab effects (other than lab F) are also shown to have similar values on the summary table, confirming what we saw on the posterior density plot. This suggests that there are no difference in Lab effects.

f. Briefly compare your results from parts (c) and (e). What are the main similarities, and what are the main differences, between the two approaches? [3 points]

Results are close to those found by lme4 analysis (but not exact). Both of the analyses found the intercept to be close to 87 (86.99 and 87.04, respectively). However, the first regression analysis in (c) does not tell us that much about how significant the lab effect is, and the difference between the strengths of each of the lab effects. Although we found that 0 is included in the lower bound of the confidence interval for the lab effect, with a pretty large confidence interval range, more information can be extracted using the INLA method. We further investigated the relationship between the lab effects, and found that most of the lab effects have similar means and standard deviations as well as similar posterior densities. This allowed us to analyse the differences between lab effects in more detail.

g. Now repeat parts (c) and (e) removing the outlier. There is no need to repeat every part of the analysis, but summarize the most important ways in which the analysis changes when the outlier is omitted. [4 points]

```
rromit<- rr[c(-1),]
reromit<-lmer(Strength~1+(1|Lab),rr,subset=c(-1))
#Part (c) analysis:
summary(reromit)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Strength ~ 1 + (1 | Lab)
##    Data: rr
##  Subset: c(-1)
##
## REML criterion at convergence: 313
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.00691 -0.40113 -0.01856  0.43886  2.60854
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Lab      (Intercept) 45.63    6.755
##  Residual             74.27    8.618
## Number of obs: 43, groups:  Lab, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   85.558      2.893   29.57
```

```
confint(reromit)
```

```
## Computing profile confidence intervals ...
```

```
##                 2.5 %   97.5 %
## .sig01       2.838963 12.78092
## .sigma       6.956049 11.03986
## (Intercept) 79.427454 91.53068
```

```
# The 95% confidence interval for the Lab effect is [2.84,12.78] and [6.96,11.04] for  the residual effect.

#Part (e) analysis:
formula=Strength~f(Lab,model="iid")
result=inla(formula,family="gaussian",data=rromit)
summary(result)
```

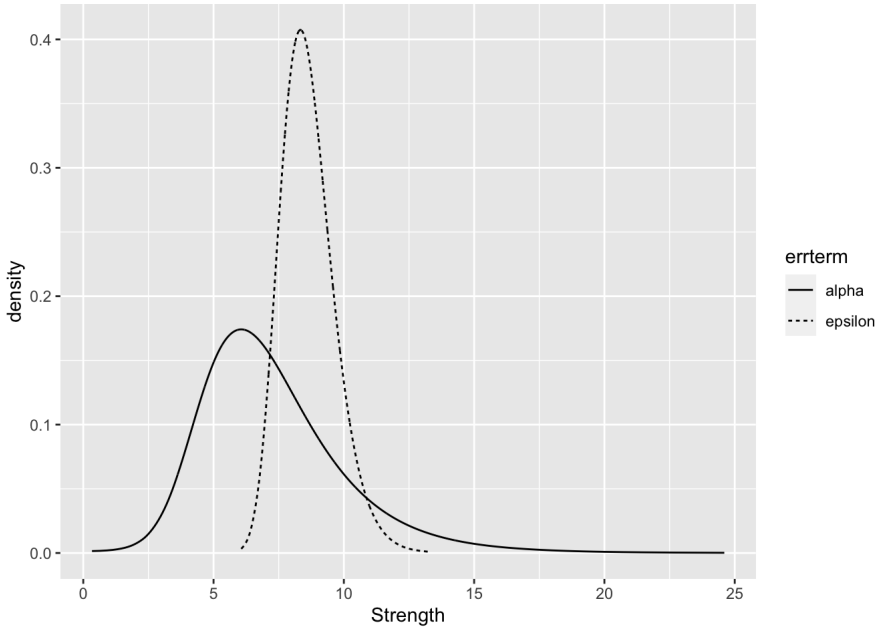
```
##
## Call:
##    "inla(formula = formula, family = \"gaussian\", data = rromit)"
## Time used:
##     Pre = 1.92, Running = 0.267, Post = 0.107, Total = 2.29
## Fixed effects:
##               mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) 86.593 1.655     83.328   86.593     89.854 86.593   0
##
## Random effects:
##   Name     Model
##     Lab IID model
##
## Model hyperparameters:
##                                           mean       sd 0.025quant 0.5quant
## Precision for the Gaussian observations 9.00e-03 2.00e-03      0.006 9.00e-03
## Precision for Lab                       1.86e+04 1.82e+04   1263.717 1.32e+04
##                                         0.975quant     mode
## Precision for the Gaussian observations    1.3e-02    0.008
## Precision for Lab                          6.7e+04 3450.384
##
## Expected number of effective parameters(stdev): 1.00(0.001)
## Number of equivalent replicates : 42.99
##
## Marginal log-Likelihood:  -177.01
```

```
sdres <- sd(rromit$Strength)
pcprior <- list(prec = list(prior="pc.prec", param = c(3*sdres,0.01)))
formula <- Strength ~ f(Lab, model="iid", hyper = pcprior)
result <- inla(formula, family="gaussian", data=rromit)
result <- inla.hyperpar(result)
summary(result)
```

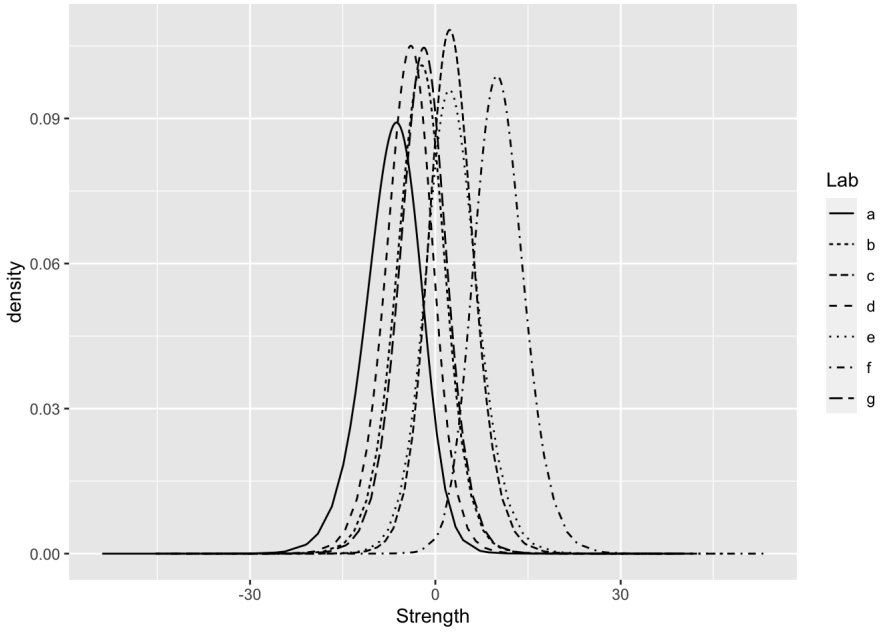
```
##
## Call:
##    "inla(formula = formula, family = \"gaussian\", data = rromit)"
## Time used:
##     Pre = 1.37, Running = 0.245, Post = 0.133, Total = 1.75
## Fixed effects:
##              mean    sd 0.025quant 0.5quant 0.975quant  mode kld
## (Intercept) 85.59 3.266     78.913   85.631     92.056 85.71   0
##
## Random effects:
##   Name     Model
##     Lab IID model
##
## Model hyperparameters:
##                                          mean     sd 0.025quant 0.5quant
## Precision for the Gaussian observations 0.014  0.003      0.008    0.014
## Precision for Lab                       0.336 66.611      0.005    0.022
##                                         0.975quant  mode
## Precision for the Gaussian observations      0.021 0.013
## Precision for Lab                            0.104 0.013
##
## Expected number of effective parameters(stdev): 5.51(0.902)
## Number of equivalent replicates : 7.80
##
## Marginal log-Likelihood:  -172.31
```

```
#Precision for Lab (sigma alpha): 0.336
#Posterior density: 1.725

#(i). A plot of posterior densities for the Lab and Residual standard deviations;
sigmaalpha <- inla.tmarginal(function(x) 1/sqrt(exp(x)),result$internal.marginals.hyperpar[[2]])
sigmaepsilon <- inla.tmarginal(function(x) 1/sqrt(exp(x)),result$internal.marginals.hyperpar[[1]])
ddf <- data.frame(rbind(sigmaalpha,sigmaepsilon),errterm=gl(2,2048,labels = c("alpha","epsilon")))
ggplot(ddf, aes(x,y, linetype=errterm))+geom_line()+xlab("Strength")+ylab("density")
```




```
#sigma alpha is not as concentrated at 0.
#(ii). A plot of posterior densities for the 7 Lab effects
rdf <- do.call(rbind.data.frame, result$marginals.random$Lab)
rdf <- cbind(Lab=gl(7,nrow(rdf)/7,labels=letters[1:7]),rdf)
ggplot(rdf, aes(x=x,y=y,linetype=Lab))+geom_line()+xlab("Strength")+ylab("density")
```




```
#Larger difference between the distributions, makes sense b/c the lab effects are significant in this model so th
e labs will have diff results.
#(iii). A summary table of posterior distributions for the main parameters of the models.
restab <- sapply(result$marginals.fixed, function(x) inla.zmarginal(x,silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaalpha,silent=TRUE))
restab <- cbind(restab, inla.zmarginal(sigmaepsilon,silent=TRUE))
restab <- cbind(restab, sapply(result$marginals.random$Lab,function(x) inla.zmarginal(x, silent=TRUE)))
colnames(restab) = c("mu","alpha","epsilon",levels(rromit$Lab))
data.frame(restab)
```

```
##                  mu    alpha  epsilon         A          B          C         D
## mean       85.59032 7.271665 8.626938 -6.943292   -2.4673   2.625995 -4.323909
## sd         3.266707 2.795524 1.029381  4.656451  4.251334    3.97501  4.066161
## quant0.025 78.90119 3.111744 6.896479 -16.75689 -11.15752   -5.04824 -12.71228
## quant0.25  83.61867 5.345925 7.894354 -9.889284 -5.148639 0.04553898 -6.879115
## quant0.5   85.61529  6.81029 8.524837 -6.744204 -2.406516   2.506566 -4.231529
## quant0.75  87.56855 8.677188 9.245917 -3.805515 0.2431612   5.073807 -1.701987
## quant0.975 92.03344 14.08037 10.93313  1.558056   5.78264   10.77593  3.438902
##                  E        F         G
## mean         2.728687 10.42973 -2.049869
## sd           4.452962 4.213805  4.119338
## quant0.025  -5.795382 2.632583 -10.41813
## quant0.25  -0.1969306 7.564924 -4.649185
## quant0.5     2.567796 10.23883 -2.014138
## quant0.75    5.482681 13.05308 0.5559569
## quant0.975   11.91447 19.22609  6.010757
```

```
restab
```

```
##            mu       alpha    epsilon  A         B         C          D
## mean       85.59032 7.271665 8.626938 -6.943292 -2.4673   2.625995   -4.323909
## sd         3.266707 2.795524 1.029381 4.656451  4.251334  3.97501    4.066161
## quant0.025 78.90119 3.111744 6.896479 -16.75689 -11.15752 -5.04824   -12.71228
## quant0.25  83.61867 5.345925 7.894354 -9.889284 -5.148639 0.04553898 -6.879115
## quant0.5   85.61529 6.81029  8.524837 -6.744204 -2.406516 2.506566   -4.231529
## quant0.75  87.56855 8.677188 9.245917 -3.805515 0.2431612 5.073807   -1.701987
## quant0.975 92.03344 14.08037 10.93313 1.558056  5.78264   10.77593   3.438902
##            E          F        G
## mean       2.728687   10.42973 -2.049869
## sd         4.452962   4.213805 4.119338
## quant0.025 -5.795382  2.632583 -10.41813
## quant0.25  -0.1969306 7.564924 -4.649185
## quant0.5   2.567796   10.23883 -2.014138
## quant0.75  5.482681   13.05308 0.5559569
## quant0.975 11.91447   19.22609 6.010757
```

The variance and standard deviation for the Lab effect increases from 25.51 and 5.05 to 45.63 and 6.76, respectively. For the Residual effect, the variance decreases from 124.19 to 74.27 and the standard deviation decreases from 11.14 to 8.62. Before removing the outlier, the 95% confidence interval for the Lab effect was [0,10.64] and [9.03,14.15] for the residual effect. After removing the outlier, the 95% confidence interval for the Lab effect is [2.84,12.78] and [6.96,11.04] for the residual effect. Both intervals are slightly more narrow than before omitting the outlier, especially for the Lab effect, which also no longer includes 0 in the interval. For part (i), the plot of posterior densities for the Lab and Residual

standard deviations shifted closer together, with the alpha increasing by a lot and the epsilon distribution decreasing. Sigma alphas are less concentrated around 0 as compared to before as the mean increased to 1.73, which is much further from 0. Previously, the mean was much smaller (0.013). That also falls inline with the confidence interval no longer having 0 in the interval anymore. The summary table results also show that there is a larger difference in the lab effects than previously, especially with Lab F. These exemplify that removing the outlier makes the lab effect more significantly different from each other. This is similar to what we found in part (b) when we used standard diagnostics to find that removing the outlier would make the lab effect statistically significant. Although the two analyses came to two different conclusions, I believe that it would be beneficial to examine both (instead of just one or the other) to help compare the performance of the labs. The intial analyses helped determine which lab may be the worst performer, and after removing the worst performer, it can show the smaller discrepancies between the labs for a more precise comparison of the performances.

##Question 2 *Chp 10 & 11* 2. Five varieties of barley were planted in six different fields over two years | see Table 2. The data (in a form suitable for analysis in R) a

```
r barley=read.csv('/Users/SylviaSzarka/Desktop/School/STOR 590/FINAL/barley.csv')
```

(a) Analyze the data as a fixed effects analysis of variance, treating “Yield” as the response. Are each of the Variety, Field and Year effects statistically significant?

```
r fmod<- aov(Yield~Variety+Field+Year+Variety*Field+Variety*Year,barley) summary(fmod)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F) ## Variety        4   5310    1327   2.149 0.09018 . ## Field          1   6487
```

```
r #Variety*Field p-value=0.789 & Variety*Year p-value=0.975 -> Neither interction term is significant.
```

The field and year effects are statist

(b) The experimenter is ultimately interested in differentiating different varieties of barley, whereas the Field and Year influences are random. Therefore, we would li

```
r op <- options(contrasts=c("contr.sum", "contr.poly")) options(op) #i. Year as a random effect and ignore Field; mmod1<-lmer(Yield~V
```

```
## Linear mixed model fit by REML ['lmerMod'] ## Formula: Yield ~ Variety + (1 | Year) ##    Data: barley ## ## REML criterion at cor
```

```r
#The estimate of variance components for the Year variable is 660 and the residual variance is 105. The interclass correlation is 104.63/((104.63+659.70))=0.14
```

#ii. Treat both Year and Field as separate random effects; mmod2<-lmer(Yield~Variety+(1|Year)+(1|Field),barley) summary(mmod2) ```

```
## Linear mixed model fit by REML ['lmerMod'] ## Formula: Yield ~ Variety + (1 | Year) + (1 | Field) ##    Data: barley ## ## REML cr
```

```r
#The estimate of variance components for they Field effect is 395 and Year effect is 117 and Residual effect = 294. The interclass correlation are 117/((117+294
```

#iii. Treat both Year and Field as random effects but with Year nested within Field. mmod3<-lmer(Yield~Variety+(1|Field:Year),barley) summary(mmod3) ```

```
## Linear mixed model fit by REML ['lmerMod'] ## Formula: Yield ~ Variety + (1 | Field:Year) ##    Data: barley ## ## REML criterion
```

```
r #The estimate of variance components for the Intercept is 546 and the residual variance is 171 The interclass correlation is 546/((
```

(c) For each of models (i), (ii), (iii), refit the model without Variety and perform a Kenward- Roger test for significance of the Variety effect. Why do the three models

```
r #i. Year as a random effect and ignore Field; mmodi<-lmer(Yield~1+(1|Year),barley) #testing for Variety effect: KRmodcomp(mmod1,mmo
```

```
## F-test with Kenward-Roger approximation; time: 0.19 sec ## large : Yield ~ Variety + (1 | Year) ## small : Yield ~ 1 + (1 | Year)
```

```
r #ii. Treat both Year and Field as separate random effects; mmodii<-lmer(Yield~1+(1|Year)+(1|Field),barley) KRmodcomp(mmod2,mmodii)
```

```
## F-test with Kenward-Roger approximation; time: 0.09 sec ## large : Yield ~ Variety + (1 | Year) + (1 | Field) ## small : Yield ~ 1
```

```
r #iii. Treat both Year and Field as random effects but with Year nested within Field. mmodiii<-lmer(Yield~1+(1|Field:Year),barley) K
```

```
## F-test with Kenward-Roger approximation; time: 0.04 sec ## large : Yield ~ Variety + (1 | Field:Year) ## small : Yield ~ 1 + (1 |
```

(d) Now conduct a formal test of model (i) against model (ii) against model (iii) using either a parametric bootstrap or the exactRLRT procedure. After conducting tl

```
r exactRLRT(mmod2,mmod3,mmod1)
```

```
## ##  simulated finite sample distribution of RLRT. ## ##  (p-value based on 10000 simulated values) ## ## data: ## RLRT = 44.951, p
```

```
r #p-value = 1e-04 exactRLRT(mmod1,mmod3,mmod2)
```

```
## ##  simulated finite sample distribution of RLRT. ## ##  (p-value based on 10000 simulated values) ## ## data: ## RLRT = 14.646, p
```

```
r #p-value < 2.2e-16 #mmod2 alternative exactRLRT(mmod3,mmod2,mmod1)
```

```
## ##  simulated finite sample distribution of RLRT. ## ##  (p-value based on 10000 simulated values) ## ## data: ## RLRT = 30.305, p
```

```
r #p-value < 2.2e-16 #computing the Likelikhood ratio both with and without the term you want to drop. in bootstrap standard lrt is b
```

```
## Bootstrap test; time: 31.83 sec;samples: 1000; extremes: 0; ## Requested samples: 1000 Used samples: 0 Extremes: 0 ## large : Yiel
```

```
r #P-value of 1 -> this method not designed for the situation
```

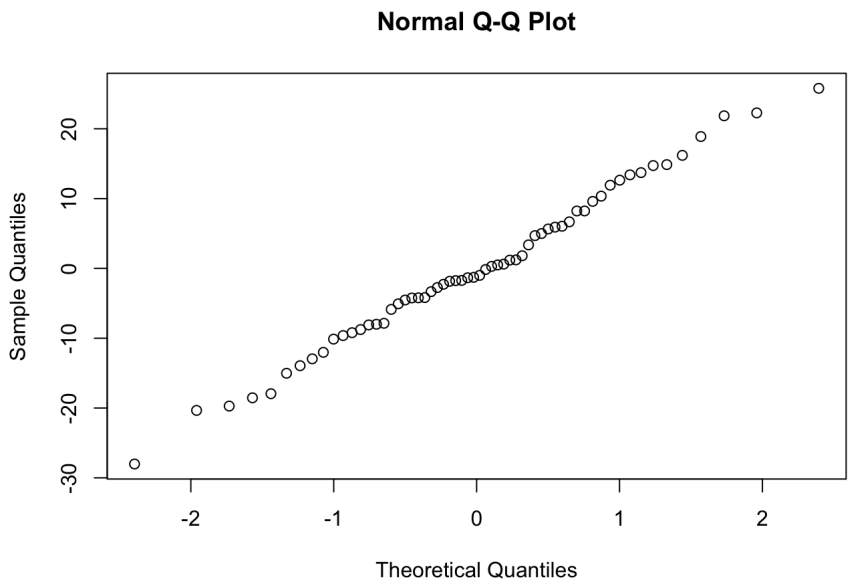
You cannot use PBmodcomp to test for a single random effect (got a p-value of 1) but

(e) Using model (iii) including the Variety effect, carry out some suitable diagnostic procedures to determine whether the model fits the data. Summarise your conc

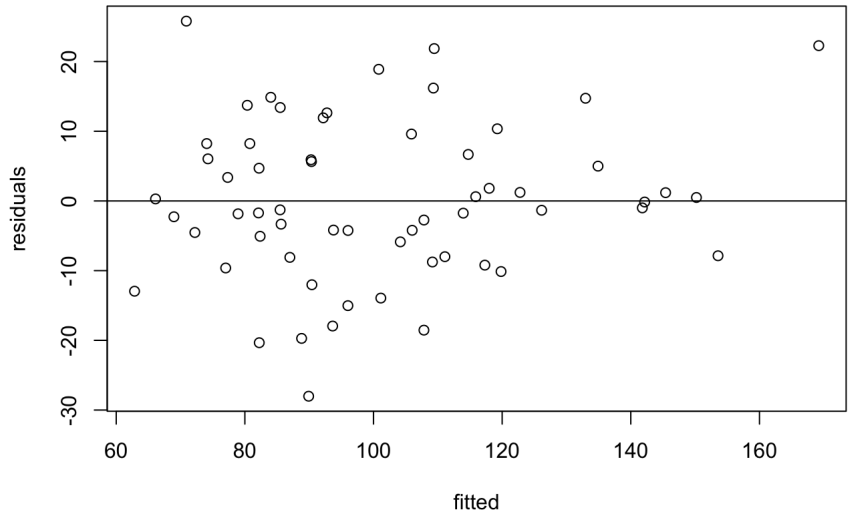
```
r dd<-fortify.merMod(mmod3) dd
```

```
##      Variety Field Year Yield   .fitted      .resid    .scresid ## 1  Manchuria     1 2010  81.0  96.01990 -15.0198953 -1.14973509
```

```
r #ggplot(dd,aes(sample=.resid))+stat_qq() qqnorm(residuals(mmod3))
```




```
r plot(fitted(mmod3),residuals(mmod3),xlab="fitted",ylab="residuals") abline(h=0)
```



The model does seem to fit the data well according t

(f) From the raw data, it looks as though Trebi is the best variety (in the sense of maximizing expected yield) and Field 2 is the best field. For the next year's crop, s

```
r #Fixed Effect = 118.1 fixef(mmod3)
```

```
##     (Intercept) VarietyPeatland VarietySvansota    VarietyTrebi   VarietyVelvet ##       94.391667        8.150000       -3.258333
```

```
r #Random effect: ranef(mmod3)
```

```
## $`Field:Year` ##        (Intercept) ## 1:2010    1.628229 ## 1:2011  -17.063460 ## 2:2010   51.020968 ## 2:2011   14.766999 ## 3:2
```

```
r #(i). Field 2 predict(mmod3, newdata=data.frame(Year='2010',Field ='2',Variety='Trebi'))
```

```
##       1 ## 169.221
```

```
r predict(mmod3, newdata=data.frame(Year='2011',Field ='2',Variety='Trebi'))
```

```
##       1 ## 132.967
```

```
r #prediction interval: predict(mmod3, newdata=data.frame(Year='2011',Field ='2',Variety='Trebi'),interval='prediction',allow.new.lev
```

```
## Warning in predict.merMod(mmod3, newdata = data.frame(Year = "2011", Field = ## "2", : unused arguments ignored
```

```
##       1 ## 132.967
```

```
```r # ========================================================================== # # Parametric bootstrap method for computing 95
```

```
group.sd <- as.data.frame(VarCorr(mmod3))$sdcor[1]resid.sd <- as.data.frame(VarCorr(mmod3))sdcor[2] pv <- numeric(1000) for(i in 1:1000){ y <- unlist(si
```
```

```
##     2.5%    97.5% ## 115.0423 217.2184
```

```r # 2.5% 97.5% # 111.7868 216.5640
```

```
# ============================================================================== # # Parametric bootstrap method for computing 95%
```

```
##      2.5%     97.5% ##  84.40538 182.44059
```

```
r ##(ii). # ============================================================================== # Prediction for a new randomly chosen fie
```

```
##        1 ## 94.39167
```

```r # 94.39167
```

```
# ============================================================================== # # Parametric bootstrap method for computing 95%
```

group.sd <- as.data.frame(VarCorr(mmod3))$sdcor[1] $resid.sd <- as.data.frame(VarCorr(mmod3))$sdcor[2] pv <- numeric(1000) for(i in 1:1000){ y <- unlist(si

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : ## Model failed to converge with max|grad| = 0.0028
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : ## Model failed to converge with max|grad| = 0.0021
```

```
r quantile(pv, c(0.025, 0.975))
```

```
##      2.5%     97.5% ##  41.67832 149.27204
```

```
r #     2.5%     97.5% # 37.69605 153.77165
```
 (i). The point estimate for the prediction of the Yield for a crop of Variety type “Trebi” in Field 2 will be 133 fo

##Question 3: ##*Chp 13* 3. A study was conducted in which 167 mothers with children were asked to provide demographic and personal information and then followed up for 28 days each. On each day, the mother was assessed for stress and a binary variable stress (0 for low stress. 1 for high stress) was recorded. The covariates involved in the study were: id = mother-child id day = study day t=(1,2,…,28) stress = maternal stress at day(t): 1=yes, 0=no married = marital status: 1=married, 0=other education = highest educational level: 1=less than high school, 2=some high school, 3=high school graduate, 4=some college, 5=college graduate employed = employment status: 1=employed, 0=unemployed chlth = child health status at baseline: 1=very poor,2=poor,3=fair,4=good,5=very good mhlth = mother health status at baseline: 1=very poor,2=poor,3=fair,4=good,5=very good race = child race: 1=non-white, 0=white csex = child gender: 1=male, 2=female housize = size of household: 1=more than 3 people, 0=2-3 people

```
stress=read.csv('/Users/SylviaSzarka/Desktop/School/STOR 590/FINAL/stress.csv')
```

a. Construct a plot in which “mean stress level” is plotted against “day,” averaging over individuals, with employed and unemployed mothers shown with different plotting symbols on the same plot. Also fit a straight line to the plot, separately for employed and unemployed mothers. You should observe that both groups show decreased stress levels over time, but that the relationship is not the same for the employed and unemployed mothers. Describe the relationships. [7 points]

```
#summary(stress)
#33 NA values in stress..omitting the values
stress1<-na.omit(stress)
#Check: Does mother employed status change? -> No.
length(unique(stress$id))
```

```
## [1] 167
```

```
nrow(unique(stress[,c('id','employed')]))
```

```
## [1] 167
```

```
#Mean stress level per day broken down by unemployed vs. employed:
stress1 %>%
  group_by(employed,day) %>%
  summarize(meanstress=mean(stress)) %>%
  xtabs(formula=meanstress~day+employed)
```

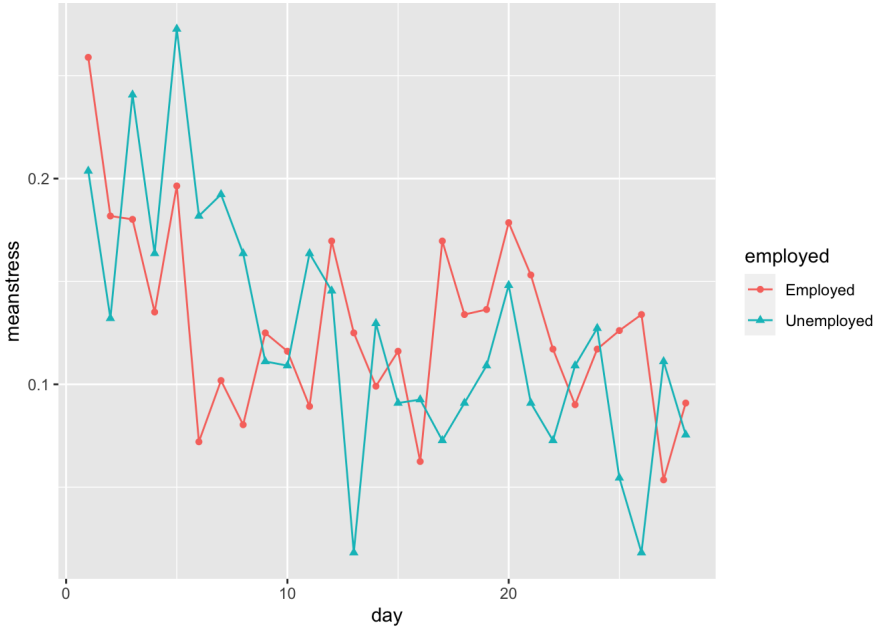
```
## `summarise()` regrouping output by 'employed' (override with `.groups` argument)
```

```
##     employed
## day           0          1
##   1  0.20370370 0.25892857
##   2  0.13207547 0.18181818
##   3  0.24074074 0.18018018
##   4  0.16363636 0.13513514
##   5  0.27272727 0.19642857
##   6  0.18181818 0.07207207
##   7  0.19230769 0.10185185
##   8  0.16363636 0.08035714
##   9  0.11111111 0.12500000
##   10 0.10909091 0.11607143
##   11 0.16363636 0.08928571
##   12 0.14545455 0.16964286
##   13 0.01818182 0.12500000
##   14 0.12962963 0.09909910
##   15 0.09090909 0.11607143
##   16 0.09259259 0.06250000
##   17 0.07272727 0.16964286
##   18 0.09090909 0.13392857
##   19 0.10909091 0.13636364
##   20 0.14814815 0.17857143
##   21 0.09090909 0.15315315
##   22 0.07272727 0.11711712
##   23 0.10909091 0.09009009
##   24 0.12727273 0.11711712
##   25 0.05454545 0.12612613
##   26 0.01818182 0.13392857
##   27 0.11111111 0.05357143
##   28 0.07547170 0.09090909
```

```
#Grouping
df <- stress1 %>%
    group_by(day,employed) %>%
    summarise(meanstress = mean(stress), emp=mean(employed))
```

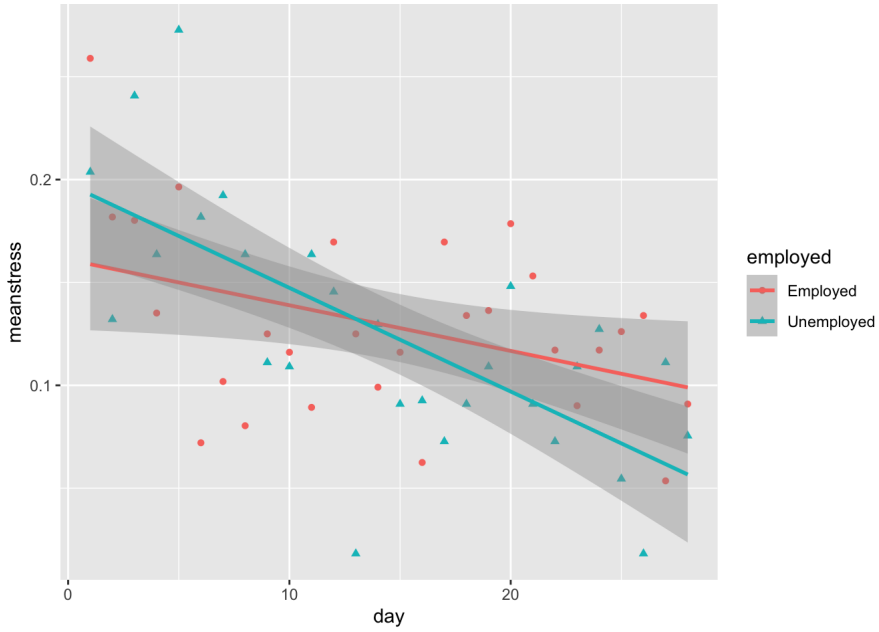
```
## `summarise()` regrouping output by 'day' (override with `.groups` argument)
```

```
df$employed <- ifelse(df$employed == 1, "Employed", "Unemployed")
#Averaging based on mean stress levels over the 28 days for unemployed & employed separately
ggplot(df,aes(x=day,y=meanstress,shape=employed,color=employed)) +geom_point()+geom_line()
```



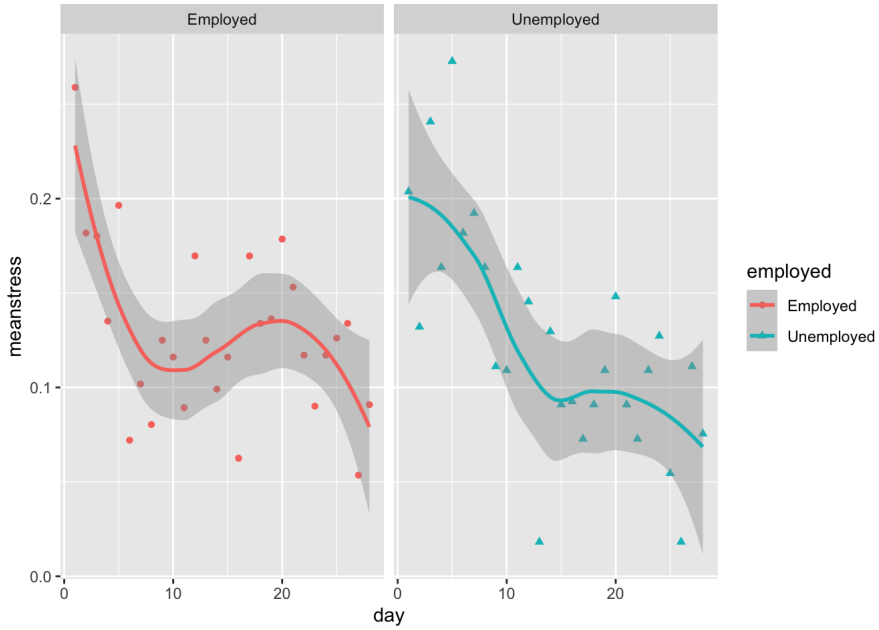

```
ggplot(df,aes(x=day,y=meanstress,shape=employed,color=employed)) +geom_point()+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(df,aes(x=day,y=meanstress,shape=employed,color=employed))+geom_point()+geom_smooth()+facet_wrap(~employed)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Both of the plots shows that both groups show decreased stress levels over time. The fitted lines in the second plot show that for unemployed mothers, the stress level starts pretty high but then sharply decreases mid-month, making the slope for that group a lot larger. The employed mothers have a much smaller slope, showing a more steady decline in stress over the 28-day period.

b. The other variables in the analysis are likely to be correlated with the mother's employment status, and therefore could be confounders to the relationship you observed in (a). With this in mind, fit a GLMM to the whole of the data, including all of day, married, factor(education), employed, factor(chlth), factor(mhlth), race, csex and housize as covariates, but also including a day:employed interaction term. Do this using:

i. PQL method,
ii. glmer method, iii.GEE method with corstr='ar1'
iii. GEE method with corstr='exchangeable' Compare these methods, with particular focus on the statistical significance of the *day:employed* interaction term. Which method or methods do you think work best for this problem? [12 points] *ctsib example*

```
#i. PQL method:
modpql1=glmmPQL(stress~I(day*employed)+day+married+factor(education)+employed+factor(chlth)+factor(mhlth)+race+cs
ex+housize,random=~1|id,family=binomial,data=stress1)
```

```
## iteration 1
```

```
## iteration 2
```

```
## iteration 3
```

```
## iteration 4
```

```
  summary(modpql1)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: stress1
##    AIC BIC logLik
##     NA  NA     NA
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept)  Residual
## StdDev:   0.9522576 0.8823728
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~invwt
## Fixed effects: stress ~ I(day * employed) + day + married + factor(education) +      employed + factor(chlth)
+ factor(mhlth) + race + csex +      housize
##                         Value Std.Error   DF   t-value p-value
## (Intercept)         1.5912691 1.5993840 4474  0.994926  0.3198
## I(day * employed)   0.0328808 0.0113572 4474  2.895147  0.0038
## day                -0.0549387 0.0095172 4474 -5.772594  0.0000
## married             0.3790411 0.1995834  149  1.899161  0.0595
## factor(education)2  0.2455075 0.3641478  149  0.674197  0.5012
## factor(education)3  0.6806534 0.3808130  149  1.787369  0.0759
## factor(education)4  0.6520929 0.4354808  149  1.497409  0.1364
## factor(education)5 -0.6539575 0.9194355  149 -0.711260  0.4780
## employed           -0.1135465 0.2628117  149 -0.432045  0.6663
## factor(chlth)2     -1.1801493 1.1239911  149 -1.049963  0.2954
## factor(chlth)3     -1.7541528 1.0833382  149 -1.619211  0.1075
## factor(chlth)4     -1.7410034 1.0774929  149 -1.615791  0.1083
## factor(chlth)5     -1.9994403 1.0824232  149 -1.847189  0.0667
## factor(mhlth)2     -1.9499147 1.0964524  149 -1.778385  0.0774
## factor(mhlth)3     -1.6994309 1.0782596  149 -1.576087  0.1171
## factor(mhlth)4     -2.1318434 1.0967301  149 -1.943818  0.0538
## factor(mhlth)5     -2.5817445 1.1173781  149 -2.310538  0.0222
## race                0.1159542 0.2044299  149  0.567207  0.5714
## csex                0.0905748 0.1880902  149  0.481550  0.6308
## housize            -0.5034536 0.2028475  149 -2.481932  0.0142
##  Correlation:
##                    (Intr) I(*em) day    marrid fctr(d)2 fctr(d)3 fctr(d)4
## I(day * employed)   0.057
## day                -0.072 -0.838
## married             0.113  0.002 -0.004
## factor(education)2 -0.280  0.000 -0.001 -0.090
## factor(education)3 -0.267  0.004 -0.005 -0.174  0.813
## factor(education)4 -0.232  0.005 -0.006 -0.148  0.705    0.763
## factor(education)5 -0.159 -0.005  0.006 -0.095  0.342    0.409    0.379
## employed           -0.126 -0.538  0.434  0.122  0.017    0.145    0.209
## factor(chlth)2     -0.656  0.003 -0.002 -0.060  0.063    0.054    0.048
## factor(chlth)3     -0.705  0.000  0.002 -0.128  0.058    0.041    0.036
## factor(chlth)4     -0.685  0.000  0.002 -0.122  0.061    0.034    0.039
## factor(chlth)5     -0.675 -0.001  0.003 -0.093  0.057    0.032    0.038
## factor(mhlth)2     -0.659  0.004 -0.001 -0.100  0.062    0.044    0.036
## factor(mhlth)3     -0.637  0.005 -0.002 -0.078  0.074    0.035    0.021
## factor(mhlth)4     -0.640  0.004 -0.001 -0.094  0.073    0.043    0.021
## factor(mhlth)5     -0.612  0.005 -0.001 -0.076  0.058    0.014   -0.009
## race               -0.169  0.007 -0.008 -0.276  0.057    0.231    0.147
## csex               -0.136 -0.004  0.004  0.086  0.020    0.005   -0.049
## housize            -0.027 -0.006  0.009 -0.228 -0.040   -0.022    0.042
##                    fctr(d)5 emplyd fctr(c)2 fctr(c)3 fctr(c)4 fctr(c)5 fctr(m)2
## I(day * employed)
## day
## married
## factor(education)2
## factor(education)3
## factor(education)4
## factor(education)5
## employed            0.176
## factor(chlth)2      0.059    0.043
## factor(chlth)3      0.039   -0.019  0.907
## factor(chlth)4      0.040   -0.008  0.927    0.961
## factor(chlth)5      0.038   -0.003  0.919    0.953    0.979
## factor(mhlth)2      0.007   -0.018  0.015    0.098    0.037    0.038
## factor(mhlth)3     -0.002   -0.010 -0.026    0.049   -0.016   -0.016    0.964
## factor(mhlth)4      0.008    0.002 -0.017    0.062   -0.018   -0.022    0.961
## factor(mhlth)5     -0.037   -0.027 -0.021    0.053   -0.018   -0.037    0.939
## race                0.176   -0.105  0.054    0.093    0.092    0.073    0.103
## csex                0.029    0.020  0.114    0.045    0.085    0.067   -0.135
## housize             0.139   -0.029  0.098    0.039    0.098    0.085   -0.153
##                    fctr(m)3 fctr(m)4 fctr(m)5 race   csex
## I(day * employed)
## day
## married
```

```
## factor(education)2
## factor(education)3
## factor(education)4
## factor(education)5
## employed
## factor(chlth)2
## factor(chlth)3
## factor(chlth)4
## factor(chlth)5
## factor(mhlth)2
## factor(mhlth)3
## factor(mhlth)4      0.979
## factor(mhlth)5      0.959    0.961
## race                0.068    0.086    0.069
## csex               -0.119   -0.118   -0.125   -0.150
## housize            -0.158   -0.149   -0.160    0.075  0.040
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -1.7121444 -0.4511658 -0.2923002 -0.1934910  6.0212571
##
## Number of Observations: 4643
## Number of Groups: 167
```

```
  #(day*employed) interaction significant - p-value of 0.0038
  #difficult to interpret w/ all the levels for the variables

#Try without factoring variables (easier to interpret)
modpql2=glmmPQL(stress~I(day*employed)+day+married+education+employed+chlth+mhlth+race+csex+housize,random=~1|id,
family=binomial,data=stress1)
```

```
## iteration 1
```

```
## iteration 2
```

```
## iteration 3
```

```
## iteration 4
```

```
## iteration 5
```

```
  summary(modpql2)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: stress1
##    AIC BIC logLik
##     NA  NA     NA
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept)  Residual
## StdDev:    0.999418 0.8812717
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~invwt
## Fixed effects: stress ~ I(day * employed) + day + married + education + employed +      chlth + mhlth + race +
csex + housize
##                        Value Std.Error   DF   t-value p-value
## (Intercept)       -0.2086012 0.6814520 4474 -0.306113  0.7595
## I(day * employed)  0.0327910 0.0113282 4474  2.894633  0.0038
## day               -0.0548475 0.0094921 4474 -5.778257  0.0000
## married            0.3581730 0.1978851  158  1.810005  0.0722
## education          0.1922215 0.1259736  158  1.525887  0.1290
## employed          -0.1458041 0.2636557  158 -0.553009  0.5810
## chlth             -0.2449211 0.1230033  158 -1.991175  0.0482
## mhlth             -0.2808824 0.1149732  158 -2.443025  0.0157
## race               0.1498991 0.2010787  158  0.745475  0.4571
## csex               0.0359686 0.1880420  158  0.191279  0.8486
## housize           -0.4940038 0.2002213  158 -2.467289  0.0147
##  Correlation:
##                   (Intr) I(*em) day    marrid eductn emplyd chlth  mhlth
## I(day * employed)  0.143
## day               -0.172 -0.838
## married           -0.037  0.002 -0.004
## education         -0.479  0.004 -0.006 -0.172
## employed          -0.378 -0.535  0.432  0.111  0.306
## chlth             -0.444 -0.009  0.011 -0.018 -0.026 -0.039
## mhlth             -0.180  0.004 -0.002  0.024 -0.143  0.026 -0.462
## race              -0.181  0.008 -0.008 -0.224  0.252 -0.112  0.031 -0.038
## csex              -0.332 -0.004  0.004  0.089 -0.063 -0.007 -0.018  0.001
## housize           -0.234 -0.005  0.008 -0.262  0.102 -0.069  0.069 -0.015
##                   race   csex
## I(day * employed)
## day
## married
## education
## employed
## chlth
## mhlth
## race
## csex              -0.161
## housize            0.086 -0.018
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -1.7116363 -0.4529547 -0.2935365 -0.1932410  5.8078705
##
## Number of Observations: 4643
## Number of Groups: 167
```

```
  #(day*employed) interaction still significant w/ same pvalue as before

#ii. glmer method* *UNABLE TO ADD OTHER VARIABLES*
  #Uses standard likelihood based methods to construct a chi-squared test --> view w/ skepticism bc of the shortc
omings/failures of chi-square approximations.
    #a) Using default
    modgha=glmer(stress~day+married+I(day*employed)+(1+day|id),family=binomial,data=stress1)
    summary(modgha)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: stress ~ day + married + I(day * employed) + (1 + day | id)
##    Data: stress1
##
##      AIC      BIC   logLik deviance df.resid
##   3237.3   3282.4  -1611.7   3223.3     4636
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1601 -0.3921 -0.2526 -0.1561  4.7790
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 1.185425 1.08877
##         day         0.002553 0.05053  -0.17
## Number of obs: 4643, groups:  id, 167
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.93570    0.18435 -10.500  < 2e-16 ***
## day               -0.06630    0.01406  -4.716  2.4e-06 ***
## married            0.33673    0.21072   1.598   0.1100
## I(day * employed)  0.02572    0.01366   1.883   0.0597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) day    marrid
## day         -0.364
## married     -0.590 -0.029
## I(d*emplyd) -0.033 -0.708  0.085
```

```
     #as soon as I added in employed, race, csex, housize, or any factored variables it said: "Model failed to c
onverge with max|grad| = 0.0723041 (tol = 0.002, component 1) & also said Model is nearly unidentifiable."
    #b) Gauss-Hermite Approach:
    modghb=glmer(stress~day+married+I(day*employed)+(1|id),nAGQ=25,family=binomial,data=stress1)
    summary(modghb)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 25) [glmerMod]
##  Family: binomial  ( logit )
## Formula: stress ~ day + married + I(day * employed) + (1 | id)
##    Data: stress1
##
##      AIC      BIC   logLik deviance df.resid
##   3246.6   3278.8  -1618.3   3236.6     4638
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.3718 -0.3945 -0.2579 -0.1741  4.9488
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 1.344    1.159
## Number of obs: 4643, groups:  id, 167
##
## Fixed effects:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.075832   0.172430 -12.039  < 2e-16 ***
## day               -0.051383   0.009642  -5.329 9.88e-08 ***
## married            0.355104   0.213406   1.664   0.0961 .
## I(day * employed)  0.026805   0.010639   2.519   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) day    marrid
## day         -0.215
## married     -0.614 -0.076
## I(d*emplyd) -0.070 -0.789  0.095
```

```
    #dropping subject specific effects:
    modghc=glmer(stress~day+married+(1|id),nAGQ=25,family=binomial,data=stress1)
    anova(modghb,modghc)
```

```
## Data: stress1
## Models:
## modghc: stress ~ day + married + (1 | id)
## modghb: stress ~ day + married + I(day * employed) + (1 | id)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## modghc    4 3251.2 3277.0 -1621.6   3243.2
## modghb    5 3246.6 3278.8 -1618.3   3236.6 6.5459  1    0.01051 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
    #Using anova to compare with and without interaction --> significant p-value suggests that interaction term I
S significant
#Do we need to consider changing the factor variables with multiple levels to just binary?
#iii. GEE method with corstr='ar1'
modgeep1=geeglm(stress~day+married+factor(education)+employed+factor(chlth)+factor(mhlth)+race+csex+housize+I(day
*employed),id=id,corstr='ar1',scale.fix=T,data=stress1,family=binomial)
  summary(modgeep1)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + factor(education) +
##     employed + factor(chlth) + factor(mhlth) + race + csex +
##     housize + I(day * employed), family = binomial, data = stress1,
##     id = id, corstr = "ar1", scale.fix = T)
##
##  Coefficients:
##                     Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)          1.41951  0.67861  4.376   0.0365 *
## day                 -0.04944  0.01087 20.696 5.38e-06 ***
## married              0.35700  0.15823  5.091   0.0241 *
## factor(education)2   0.17348  0.32984  0.277   0.5989
## factor(education)3   0.63908  0.33827  3.569   0.0589 .
## factor(education)4   0.64566  0.39966  2.610   0.1062
## factor(education)5  -0.70083  1.09421  0.410   0.5219
## employed            -0.10264  0.25162  0.166   0.6833
## factor(chlth)2      -0.95997  0.41769  5.282   0.0215 *
## factor(chlth)3      -1.48236  0.32099 21.326 3.87e-06 ***
## factor(chlth)4      -1.40631  0.34139 16.969 3.80e-05 ***
## factor(chlth)5      -1.76493  0.34613 25.999 3.42e-07 ***
## factor(mhlth)2      -1.70022  0.34047 24.938 5.92e-07 ***
## factor(mhlth)3      -1.60033  0.32549 24.174 8.80e-07 ***
## factor(mhlth)4      -1.96804  0.34523 32.498 1.19e-08 ***
## factor(mhlth)5      -2.47003  0.39994 38.143 6.57e-10 ***
## race                 0.13486  0.17615  0.586   0.4439
## csex                 0.03678  0.16013  0.053   0.8183
## housize             -0.41723  0.16905  6.091   0.0136 *
## I(day * employed)    0.02725  0.01366  3.981   0.0460 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.1784 0.02855
## Number of clusters:   167  Maximum cluster size: 28
```

```
  #(day*employed) interaction technically significant but not highly significant; pval=0.046

#iv. GEE method with corstr='exchangeable'
modgeep2=geeglm(stress~day+married+factor(education)+employed+factor(chlth)+factor(mhlth)+race+csex+housize+I(day
*employed),id=id,corstr='exchangeable',scale.fix=T,data=stress1,family=binomial)
  summary(modgeep2)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + factor(education) +
##     employed + factor(chlth) + factor(mhlth) + race + csex +
##     housize + I(day * employed), family = binomial, data = stress1,
##     id = id, corstr = "exchangeable", scale.fix = T)
##
##  Coefficients:
##                    Estimate Std.err  Wald Pr(>|W|)
## (Intercept)          1.4281  0.6775  4.44    0.035 *
## day                 -0.0496  0.0108 21.07  4.4e-06 ***
## married              0.3591  0.1608  4.99    0.026 *
## factor(education)2   0.1600  0.3221  0.25    0.619
## factor(education)3   0.6022  0.3296  3.34    0.068 .
## factor(education)4   0.5708  0.3913  2.13    0.145
## factor(education)5  -0.6241  1.0964  0.32    0.569
## employed            -0.1424  0.2487  0.33    0.567
## factor(chlth)2      -0.9888  0.4095  5.83    0.016 *
## factor(chlth)3      -1.4648  0.3258 20.21  6.9e-06 ***
## factor(chlth)4      -1.4002  0.3468 16.30  5.4e-05 ***
## factor(chlth)5      -1.7842  0.3481 26.28  3.0e-07 ***
## factor(mhlth)2      -1.6822  0.3445 23.85  1.0e-06 ***
## factor(mhlth)3      -1.5897  0.3248 23.96  9.8e-07 ***
## factor(mhlth)4      -1.9005  0.3394 31.35  2.2e-08 ***
## factor(mhlth)5      -2.4005  0.4048 35.16  3.0e-09 ***
## race                 0.1290  0.1789  0.52    0.471
## csex                 0.0361  0.1617  0.05    0.823
## housize             -0.4124  0.1711  5.81    0.016 *
## I(day * employed)    0.0290  0.0136  4.54    0.033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0738  0.0115
## Number of clusters:   167  Maximum cluster size: 28
```

```
#(day*employed) interaction significant with p-value = 0.033
#but is difficult to interpret b/c of the multiple levels of all of the factors.
```

When attempting to use the glmer method in (ii) on all of the variables, an error message was produced with the warning “Model failed to converge..Model is nearly unidentifiable.” This could be due to the difficulty with most of the scores being reduced to binary responses, causing a lack of variability in the data to fit all the other effects simultaneously. That also makes it difficult (or impossible) to compare to the other variables that have multiple levels (0,1,2,…5). Dropping subject specific effects helped to make the model more understandable, as we wanted to know more about the variability in the population, not just the individuals. The first and last model gave similar results & a similar p-value for the interaction term, so it provides some solace in level of consistency. The first model however, had a more difficult time interpreting the factored variables and the GEE model handled these much better with a more simple and understandable output; thus we choose the last GEE method with corstr=‘exchangeable’ in (iv).

c. For your preferred method in (b), investigate whether any of the terms may be dropped from the model, and whether they affect the day:employed interaction. Which model do you choose overall as best? [10 points]

```
summary(modgeep2)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + factor(education) +
##     employed + factor(chlth) + factor(mhlth) + race + csex +
##     housize + I(day * employed), family = binomial, data = stress1,
##     id = id, corstr = "exchangeable", scale.fix = T)
##
##  Coefficients:
##                    Estimate Std.err  Wald Pr(>|W|)
## (Intercept)          1.4281  0.6775  4.44    0.035 *
## day                 -0.0496  0.0108 21.07  4.4e-06 ***
## married              0.3591  0.1608  4.99    0.026 *
## factor(education)2   0.1600  0.3221  0.25    0.619
## factor(education)3   0.6022  0.3296  3.34    0.068 .
## factor(education)4   0.5708  0.3913  2.13    0.145
## factor(education)5  -0.6241  1.0964  0.32    0.569
## employed            -0.1424  0.2487  0.33    0.567
## factor(chlth)2      -0.9888  0.4095  5.83    0.016 *
## factor(chlth)3      -1.4648  0.3258 20.21  6.9e-06 ***
## factor(chlth)4      -1.4002  0.3468 16.30  5.4e-05 ***
## factor(chlth)5      -1.7842  0.3481 26.28  3.0e-07 ***
## factor(mhlth)2      -1.6822  0.3445 23.85  1.0e-06 ***
## factor(mhlth)3      -1.5897  0.3248 23.96  9.8e-07 ***
## factor(mhlth)4      -1.9005  0.3394 31.35  2.2e-08 ***
## factor(mhlth)5      -2.4005  0.4048 35.16  3.0e-09 ***
## race                 0.1290  0.1789  0.52    0.471
## csex                 0.0361  0.1617  0.05    0.823
## housize             -0.4124  0.1711  5.81    0.016 *
## I(day * employed)    0.0290  0.0136  4.54    0.033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0738  0.0115
## Number of clusters:   167  Maximum cluster size: 28
```

```
#Drop least significant first - csex:
modgeep3=geeglm(stress~day+married+employed+factor(education)+factor(chlth)+factor(mhlth)+race+housize+I(day*empl
oyed),id=id,corstr='exchangeable',scale.fix=T,data=stress1,family=binomial)
summary(modgeep3)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + employed + factor(education) +
##     factor(chlth) + factor(mhlth) + race + housize + I(day *
##     employed), family = binomial, data = stress1, id = id, corstr = "exchangeable",
##     scale.fix = T)
##
##  Coefficients:
##                    Estimate Std.err  Wald Pr(>|W|)
## (Intercept)          1.4688  0.6283  5.46    0.019 *
## day                 -0.0497  0.0108 21.10  4.4e-06 ***
## married              0.3564  0.1610  4.90    0.027 *
## employed            -0.1448  0.2476  0.34    0.559
## factor(education)2   0.1592  0.3236  0.24    0.623
## factor(education)3   0.6044  0.3316  3.32    0.068 .
## factor(education)4   0.5762  0.3905  2.18    0.140
## factor(education)5  -0.6235  1.0945  0.32    0.569
## factor(chlth)2      -1.0126  0.3932  6.63    0.010 *
## factor(chlth)3      -1.4734  0.3150 21.88  2.9e-06 ***
## factor(chlth)4      -1.4177  0.3274 18.75  1.5e-05 ***
## factor(chlth)5      -1.8015  0.3278 30.20  3.9e-08 ***
## factor(mhlth)2      -1.6536  0.3445 23.04  1.6e-06 ***
## factor(mhlth)3      -1.5653  0.3103 25.45  4.5e-07 ***
## factor(mhlth)4      -1.8757  0.3304 32.22  1.4e-08 ***
## factor(mhlth)5      -2.3730  0.3819 38.60  5.2e-10 ***
## race                 0.1363  0.1782  0.58    0.445
## housize             -0.4136  0.1710  5.85    0.016 *
## I(day * employed)    0.0290  0.0136  4.55    0.033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0738  0.0115
## Number of clusters:   167  Maximum cluster size: 28
```

```
#Interaction term still the same

#Try dropping Education next
modgeep4=geeglm(stress~day+married+employed+factor(chlth)+factor(mhlth)+race+housize+I(day*employed),id=id,corstr
='exchangeable',scale.fix=T,data=stress1,family=binomial)
summary(modgeep4)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + employed + factor(chlth) +
##     factor(mhlth) + race + housize + I(day * employed), family = binomial,
##     data = stress1, id = id, corstr = "exchangeable", scale.fix = T)
##
##  Coefficients:
##                   Estimate Std.err  Wald Pr(>|W|)
## (Intercept)         1.8108  0.4512 16.11  6.0e-05 ***
## day                -0.0490  0.0108 20.63  5.6e-06 ***
## married             0.4370  0.1601  7.45   0.0063 **
## employed           -0.2399  0.2438  0.97   0.3249
## factor(chlth)2     -1.0304  0.4179  6.08   0.0137 *
## factor(chlth)3     -1.4461  0.3113 21.58  3.4e-06 ***
## factor(chlth)4     -1.3763  0.3451 15.91  6.7e-05 ***
## factor(chlth)5     -1.7523  0.3404 26.50  2.6e-07 ***
## factor(mhlth)2     -1.6386  0.3447 22.60  2.0e-06 ***
## factor(mhlth)3     -1.4739  0.3012 23.95  9.9e-07 ***
## factor(mhlth)4     -1.7880  0.3175 31.72  1.8e-08 ***
## factor(mhlth)5     -2.2489  0.3750 35.96  2.0e-09 ***
## race                0.0193  0.1744  0.01   0.9120
## housize            -0.3996  0.1764  5.13   0.0235 *
## I(day * employed)   0.0284  0.0136  4.38   0.0364 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0817   0.012
## Number of clusters:   167  Maximum cluster size: 28
```

```
#drop race
modgeep5=geeglm(stress~day+married+employed+factor(chlth)+factor(mhlth)+housize+I(day*employed),id=id,corstr='exc
hangeable',scale.fix=T,data=stress1,family=binomial)
summary(modgeep5)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + employed + factor(chlth) +
##     factor(mhlth) + housize + I(day * employed), family = binomial,
##     data = stress1, id = id, corstr = "exchangeable", scale.fix = T)
##
##  Coefficients:
##                   Estimate Std.err  Wald Pr(>|W|)
## (Intercept)         1.8370  0.3810 23.24  1.4e-06 ***
## day                -0.0490  0.0108 20.61  5.6e-06 ***
## married             0.4415  0.1590  7.71   0.0055 **
## employed           -0.2345  0.2409  0.95   0.3303
## factor(chlth)2     -1.0382  0.4107  6.39   0.0115 *
## factor(chlth)3     -1.4584  0.2945 24.52  7.3e-07 ***
## factor(chlth)4     -1.3883  0.3282 17.89  2.3e-05 ***
## factor(chlth)5     -1.7615  0.3273 28.96  7.4e-08 ***
## factor(mhlth)2     -1.6502  0.3273 25.41  4.6e-07 ***
## factor(mhlth)3     -1.4824  0.2924 25.69  4.0e-07 ***
## factor(mhlth)4     -1.7987  0.3091 33.87  5.9e-09 ***
## factor(mhlth)5     -2.2598  0.3594 39.53  3.2e-10 ***
## housize            -0.4003  0.1769  5.12   0.0236 *
## I(day * employed)   0.0284  0.0136  4.37   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0817   0.012
## Number of clusters:   167  Maximum cluster size: 28
```

```
#Interaction term p-value increases slightly, still sig though

#drop employed
modgeep6=geeglm(stress~day+married+factor(chlth)+factor(mhlth)+housize+I(day*employed),id=id,corstr='exchangeabl
e',scale.fix=T,data=stress1,family=binomial)
summary(modgeep6)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + factor(chlth) + factor(mhlth) +
##     housize + I(day * employed), family = binomial, data = stress1,
##     id = id, corstr = "exchangeable", scale.fix = T)
##
##  Coefficients:
##                   Estimate  Std.err  Wald Pr(>|W|)
## (Intercept)        1.64247  0.31025 28.03  1.2e-07 ***
## day               -0.04422  0.00978 20.46  6.1e-06 ***
## married            0.46719  0.16200  8.32   0.0039 **
## factor(chlth)2    -0.99349  0.41853  5.63   0.0176 *
## factor(chlth)3    -1.46469  0.29778 24.19  8.7e-07 ***
## factor(chlth)4    -1.38024  0.32734 17.78  2.5e-05 ***
## factor(chlth)5    -1.75162  0.32754 28.60  8.9e-08 ***
## factor(mhlth)2    -1.63704  0.32479 25.41  4.6e-07 ***
## factor(mhlth)3    -1.45134  0.28941 25.15  5.3e-07 ***
## factor(mhlth)4    -1.75802  0.30374 33.50  7.1e-09 ***
## factor(mhlth)5    -2.24190  0.35969 38.85  4.6e-10 ***
## housize           -0.41901  0.17541  5.71   0.0169 *
## I(day * employed)  0.02109  0.01092  3.73   0.0534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0831  0.0124
## Number of clusters:   167  Maximum cluster size: 28
```

```
#Interaction term no longer significant
#All factors except for day*employed interaction are significant now..


#investigate what happens if we drop the interaction?
modgeep7=geeglm(stress~day+married+factor(chlth)+factor(mhlth)+housize,id=id,corstr='exchangeable',scale.fix=T,da
ta=stress1,family=binomial)
summary(modgeep7)
```

```
##
## Call:
## geeglm(formula = stress ~ day + married + factor(chlth) + factor(mhlth) +
##     housize, family = binomial, data = stress1, id = id, corstr = "exchangeable",
##     scale.fix = T)
##
##  Coefficients:
##                Estimate  Std.err  Wald Pr(>|W|)
## (Intercept)     1.71487  0.30832 30.94  2.7e-08 ***
## day            -0.02958  0.00669 19.53  9.9e-06 ***
## married         0.41977  0.15633  7.21   0.0073 **
## factor(chlth)2 -1.04663  0.41708  6.30   0.0121 *
## factor(chlth)3 -1.45198  0.29654 23.98  9.8e-07 ***
## factor(chlth)4 -1.39655  0.32765 18.17  2.0e-05 ***
## factor(chlth)5 -1.76011  0.32611 29.13  6.8e-08 ***
## factor(mhlth)2 -1.64807  0.32984 24.97  5.8e-07 ***
## factor(mhlth)3 -1.50854  0.29141 26.80  2.3e-07 ***
## factor(mhlth)4 -1.82958  0.30893 35.07  3.2e-09 ***
## factor(mhlth)5 -2.30012  0.35765 41.36  1.3e-10 ***
## housize        -0.39545  0.17202  5.28   0.0215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.0814  0.0117
## Number of clusters:   167  Maximum cluster size: 28
```

```
anova(modgeep7,modgeep6)
```

```
## Analysis of 'Wald statistic' Table
##
## Model 1 stress ~ day + married + factor(chlth) + factor(mhlth) + housize + I(day * employed)
## Model 2 stress ~ day + married + factor(chlth) + factor(mhlth) + housize
##   Df   X2 P(>|Chi|)
## 1  1 3.73     0.053 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As suggested by Faraway, it is not suggested to base all of our inferences entirely on the glmmPQL method, as it is based on the linearized model with rather dubious assumptions, which means the results cannot be relied upon. Faraway also notes that the Bernoulli response may lead to biased estimates of regression coefficients as well when using the first method in (i). Although we initially assumed that the Gauss-Hermite approximation would be the best approach because of the accuracy, we saw that that model did not converge correctly, so we do not want to use that. We assume that the observations from the same subject (id) are going to have the same correlation. We choose GEE model based on the consistency as well as the fact that we are trying to model the data on population level. We choose the GEE model beecause The estimates for a GEE represent the effect of the predictors averaged across all individuals with the same predictor values. GEEs do not use random effects but model the correlation at the marginal or correlation level, so that is why it is the best model to use as compared to the other models. In the end, after dropping multiple variables, we also found that the interaction*day term is not significant and can be considered to be dropped.

d. State, in words, a summary of your conclusions. In particular, comment on whether the pattern of stress are different in employed compared with unemployed mothers, and how your conclusions may be affected by the other variables in the analysis. [5 points] Before dropping subject specific effects in the GEE model, our day*employed term was significant at p=0.033. However, after taking these variables out and just examining the model on a population level, There does not seem to be a large difference in employed and unemployed mothers' stress levels, as exemplified after dropping certain variables. This is most likely due to the difference after dropping most of the subject specific effects.