

## Validation set vs training set LDA training - Approach one

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
```

```
## 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)
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
library("ggpubr")
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

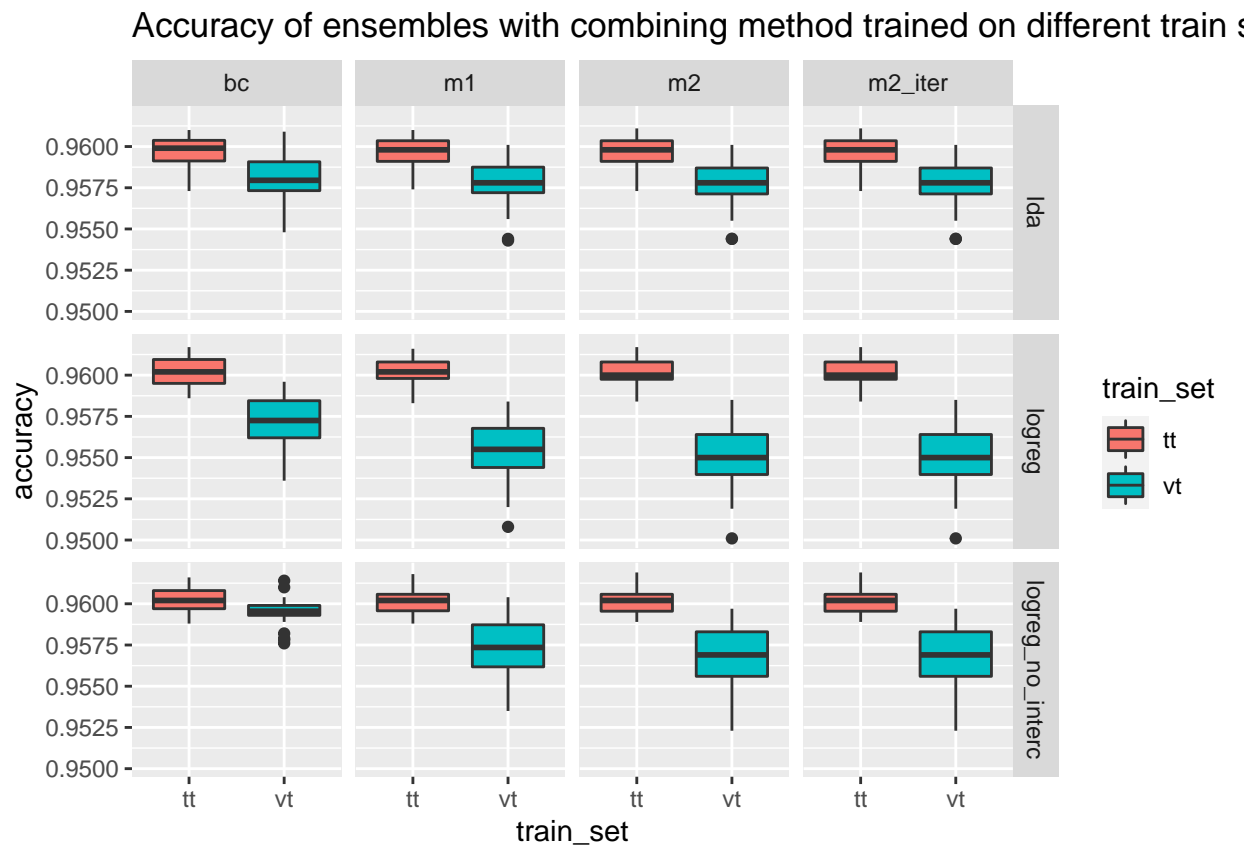
This experiment focuses on the question, whether training combining method on the same set of data as the neural networks were trained on has any adverse effects to the performance of the ensemble as opposed to training on a different set, not presented to the networks during the training. Experiment was performed with two slightly different approaches.

Approach one Experiment code is in the file `base_ensembling_experiment.py`. Experiment on CIFAR10 dataset. This experiment was performed in 30 replications. In each replication a set of 500 samples from CIFAR10 training set was randomly chosen, with each class represented equally, this set is referred to as validation set. This set was extracted from the CIFAR10 training set. Three neural networks were trained on the reduced training set. These networks were then combined using `WeightedLinearEnsemble`. For each replication, the ensemble was built twice. First linear combining method was trained on the extracted validation set and second on a randomly chosen set of 500 samples from the neural networks training set. These linear combining method training sets are referred to as vt and tt respectively.

Approach two is described and visualized in files `visualization_half_train_base_experiment_CIF10` and `visualization_half_train_base_experiment_CIF100`.

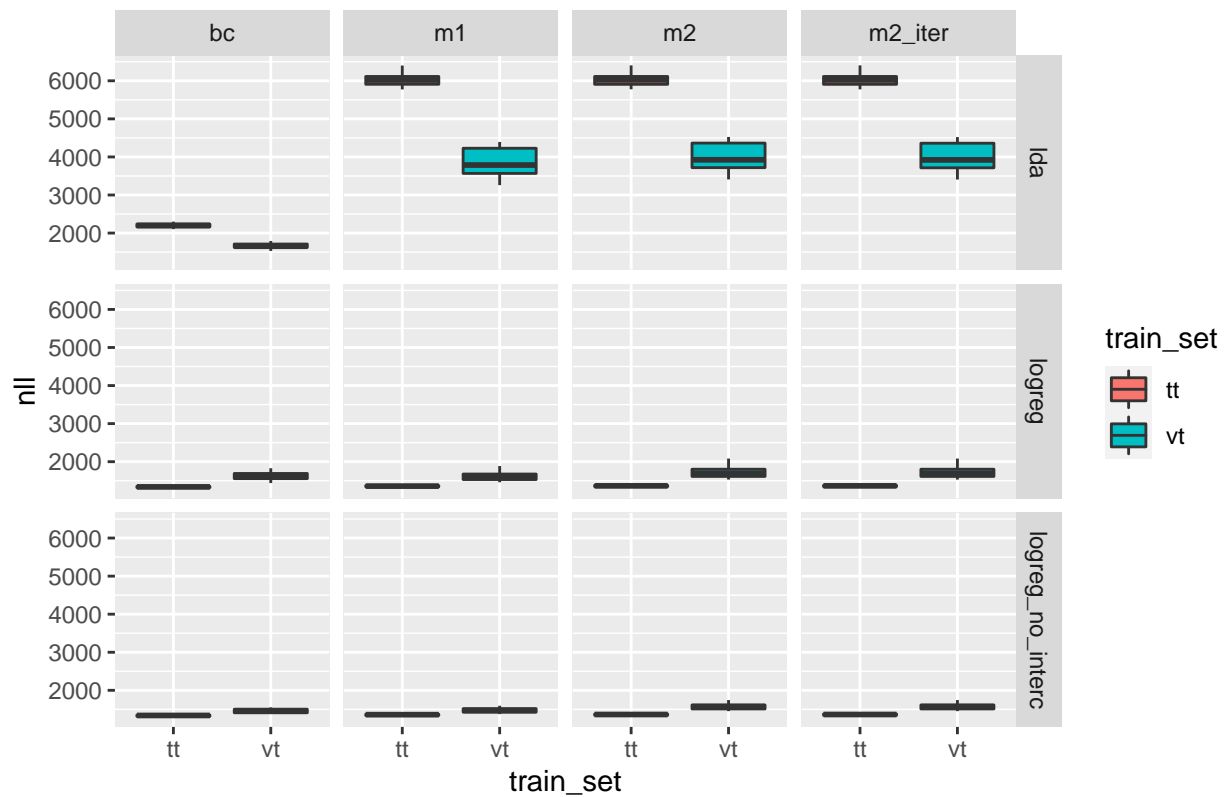
```
net_results <- read.csv("../data/data_train_val_c10/net accuracies.csv")
ens_results <- read.csv("../data/data_train_val_c10/ensemble accuracies.csv")
```

```
box_acc <- ggplot() + geom_boxplot(data=ens_results, mapping = aes(x=train_set, y=accuracy, fill=train_set),
  facet_grid(rows=vars(combining_method), cols=vars(coupling_method)) +
  ggtitle("Accuracy of ensembles with combining method trained on different train sets")
box_acc
```



```
box_nll <- ggplot() + geom_boxplot(data=ens_results, mapping = aes(x=train_set, y=nll, fill=train_set))
  facet_grid(rows=vars(combining_method), cols=vars(coupling_method)) +
  ggtitle("NLL of ensembles with combining method trained on different train sets")
box_nll
```

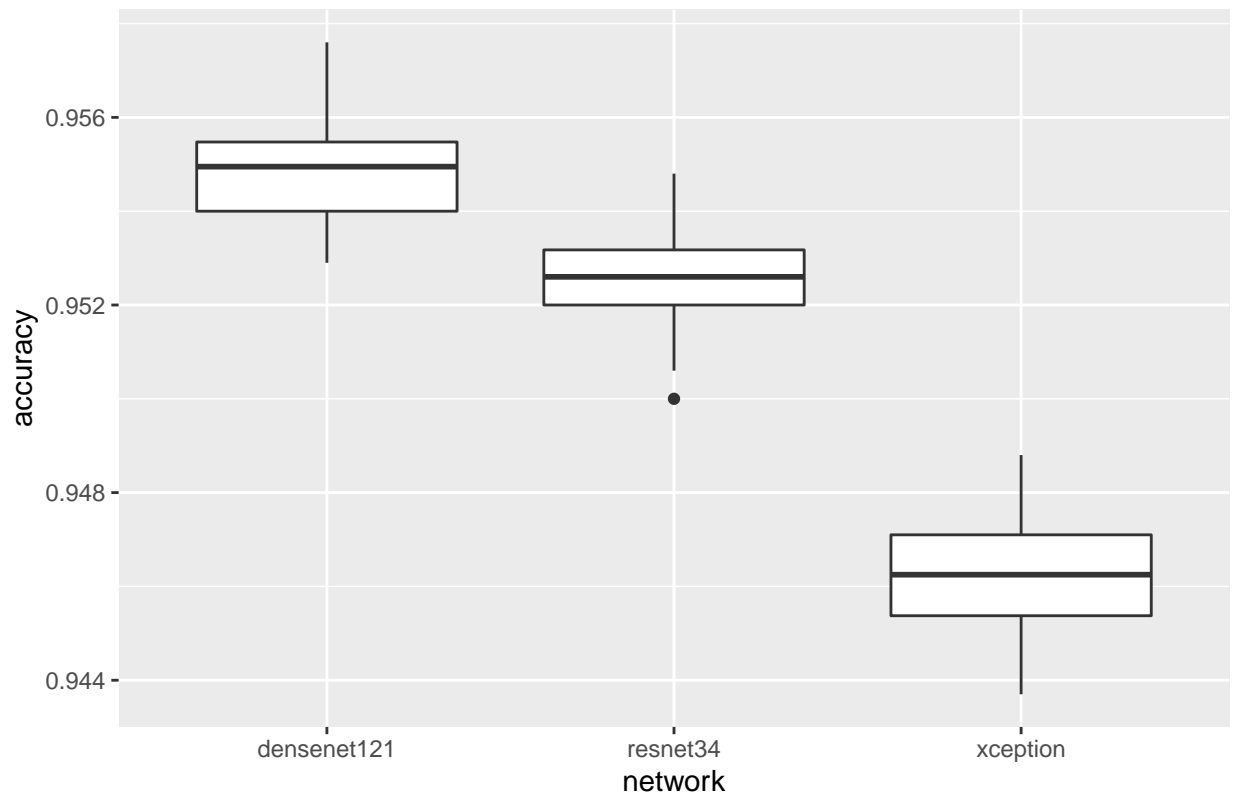
NLL of ensembles with combining method trained on different train sets



For some reason, when using lda, NLL has opposite trend to accuracy - train set with better accuracy has worse (larger) NLL. This needs some further attention. For logistic regression, the behavior is as expected.

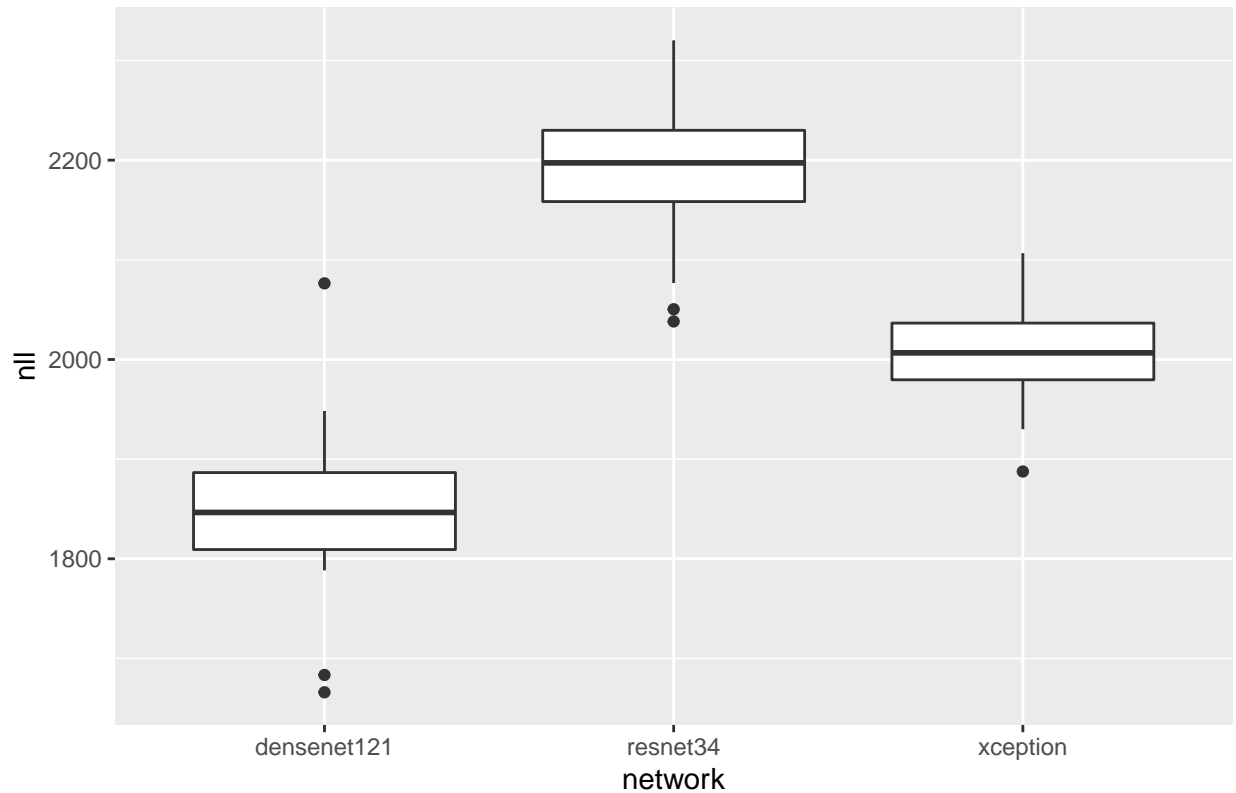
```
box_net_acc <- ggplot() + geom_boxplot(data=net_results, mapping=aes(x=network, y=accuracy)) +
  ggtitle("Accuracy of networks")
box_net_acc
```

Accuracy of networks



```
box_net_nll <- ggplot() + geom_boxplot(data=net_results, mapping=aes(x=network, y=nll)) +  
  ggtitle("NLL of networks")  
box_net_nll
```

## NLL of networks



## Accuracy

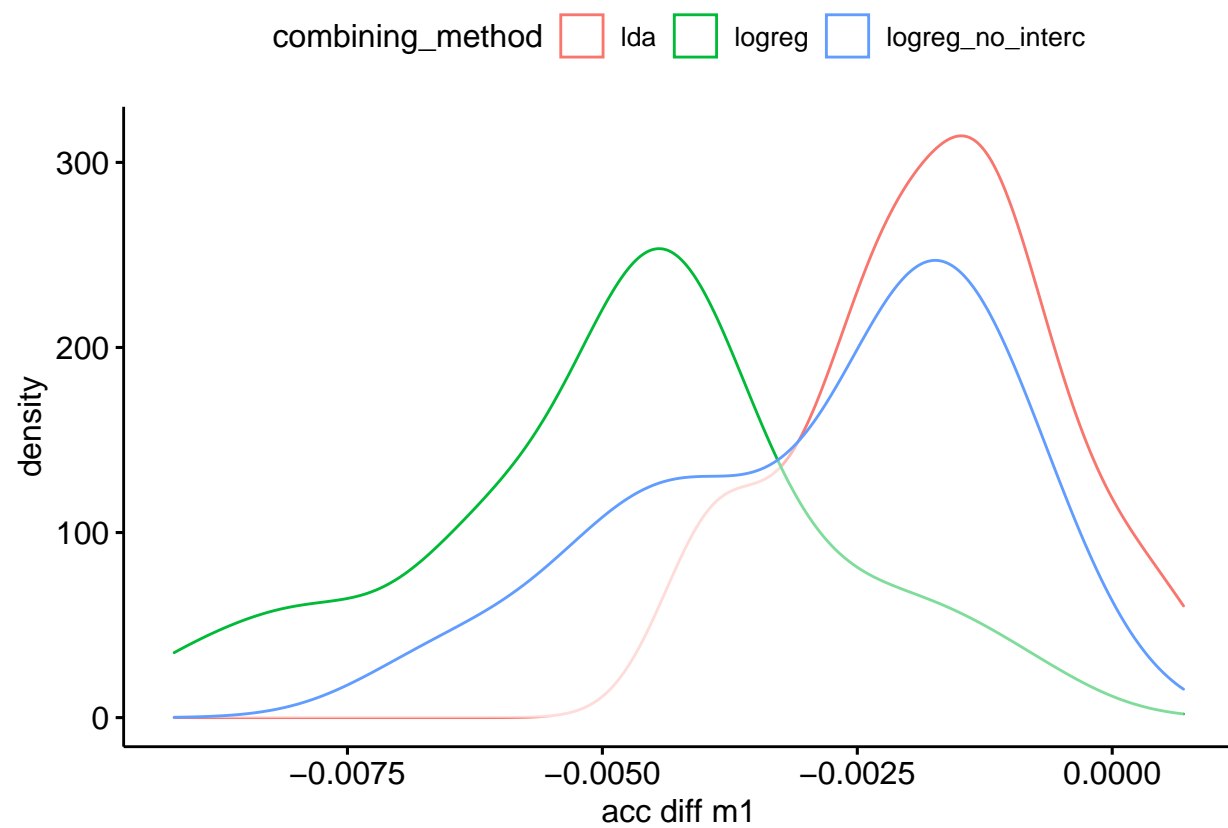
```
ens_results_wide <- pivot_wider(ens_results, names_from = c(train_set, coupling_method), values_from = c(accuracy, nll))
```

Testing statistical significance of difference for training on validation and train data Since ensembles in each replication are trained on the same set of networks, we use paired t-test.

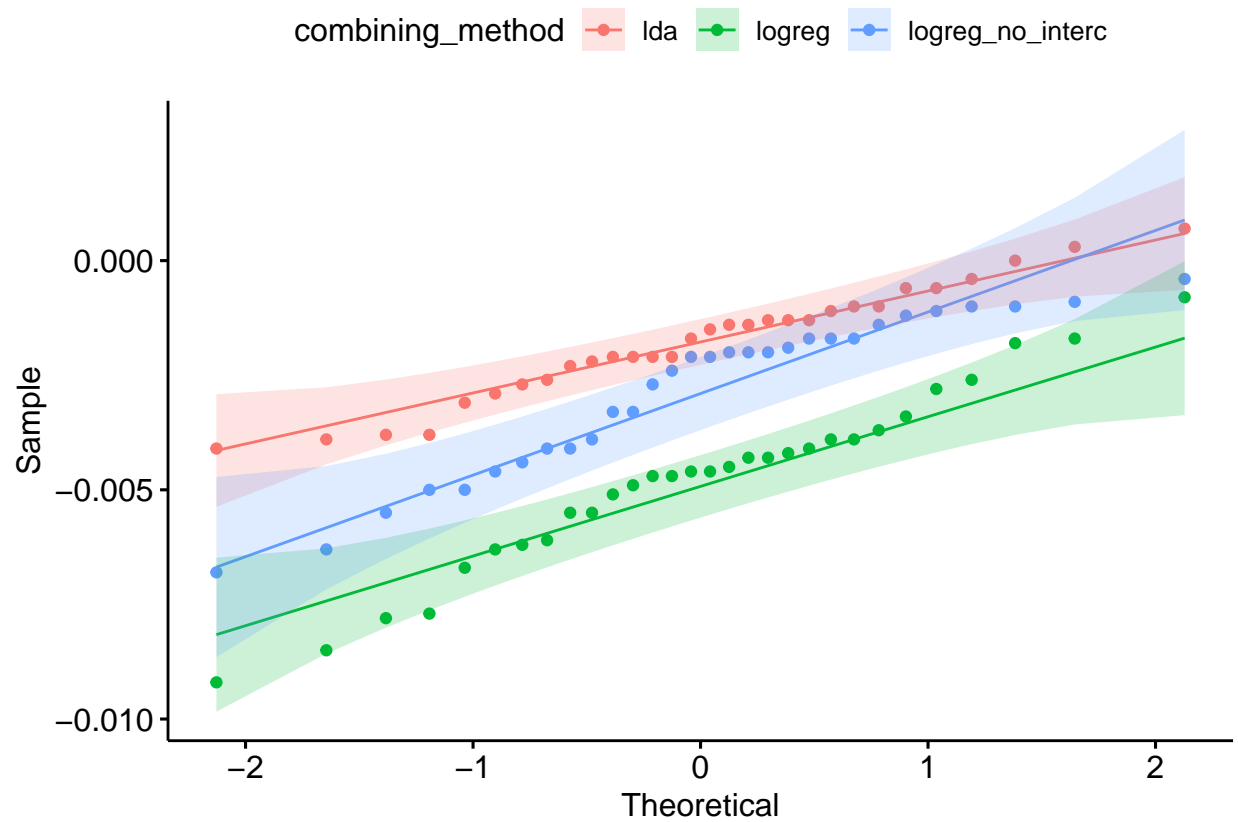
```
differences <- data.frame(ens_results_wide$combining_method,
  ens_results_wide$accuracy_vt_m1 - ens_results_wide$accuracy_tt_m1,
  ens_results_wide$accuracy_vt_m2 - ens_results_wide$accuracy_tt_m2,
  ens_results_wide$accuracy_vt_m2_iter - ens_results_wide$accuracy_tt_m2_iter,
  ens_results_wide$accuracy_vt_bc - ens_results_wide$accuracy_tt_bc,
  ens_results_wide$nll_vt_m1 - ens_results_wide$nll_tt_m1,
  ens_results_wide$nll_vt_m2 - ens_results_wide$nll_tt_m2,
  ens_results_wide$nll_vt_m2_iter - ens_results_wide$nll_tt_m2_iter,
  ens_results_wide$nll_vt_bc - ens_results_wide$nll_tt_bc)
names(differences) <- c("combining_method", "acc_diff_m1", "acc_diff_m2", "acc_diff_m2_iter", "acc_diff_m2_iter", "acc_diff_m2_iter", "acc_diff_m2_iter", "acc_diff_m2_iter")
```

Ascertaining normality of differences

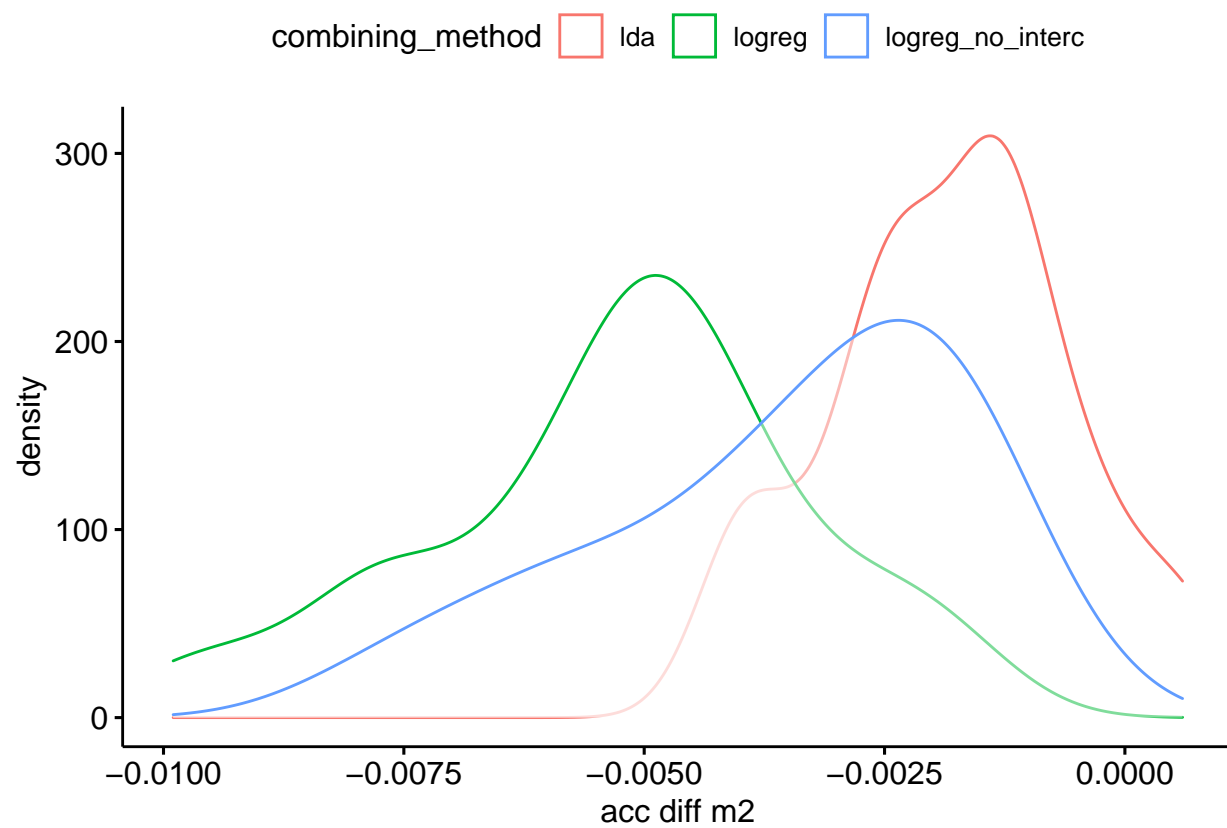
```
ggdensity(data=differences, x="acc_diff_m1", color="combining_method", xlab="acc diff m1")
```



```
ggqqplot(differences, x="acc_diff_m1", color="combining_method")
```

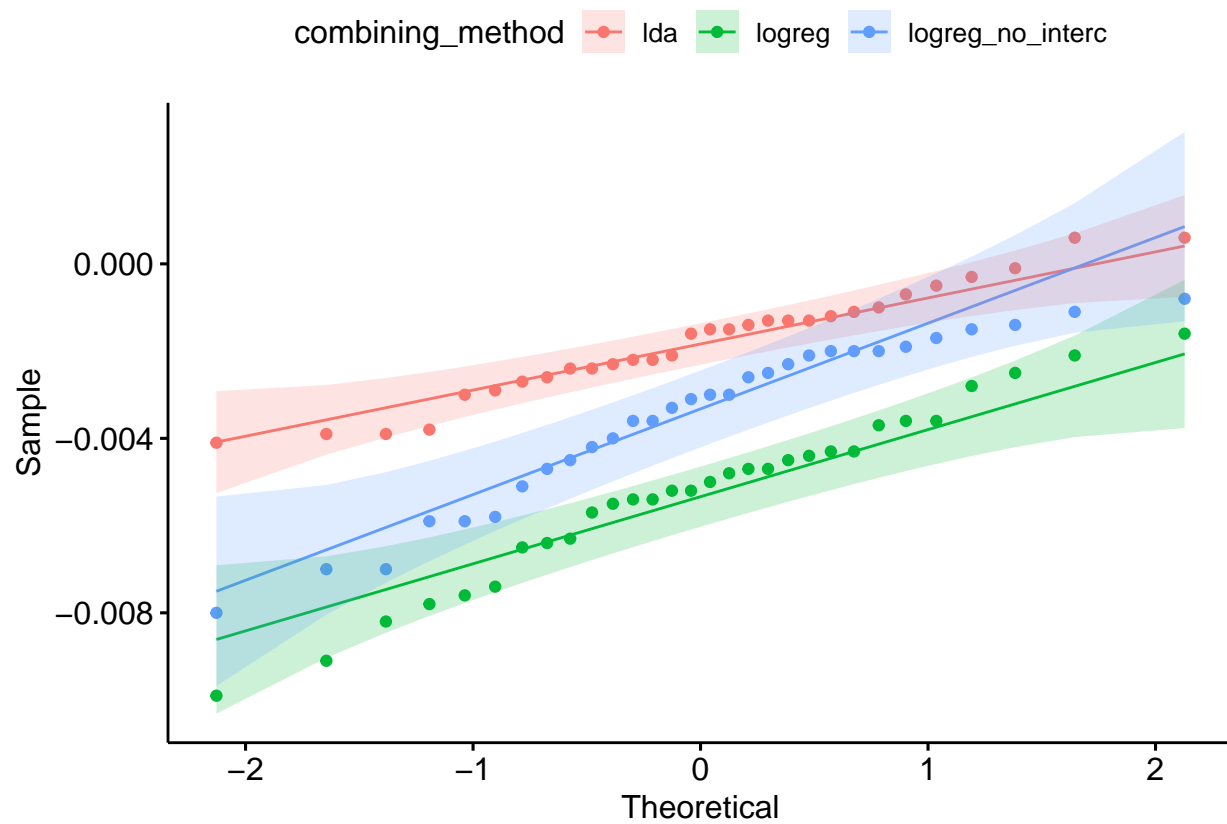


```
ggdensity(data=differences, x="acc_diff_m2", color="combining_method", xlab="acc diff m2")
```

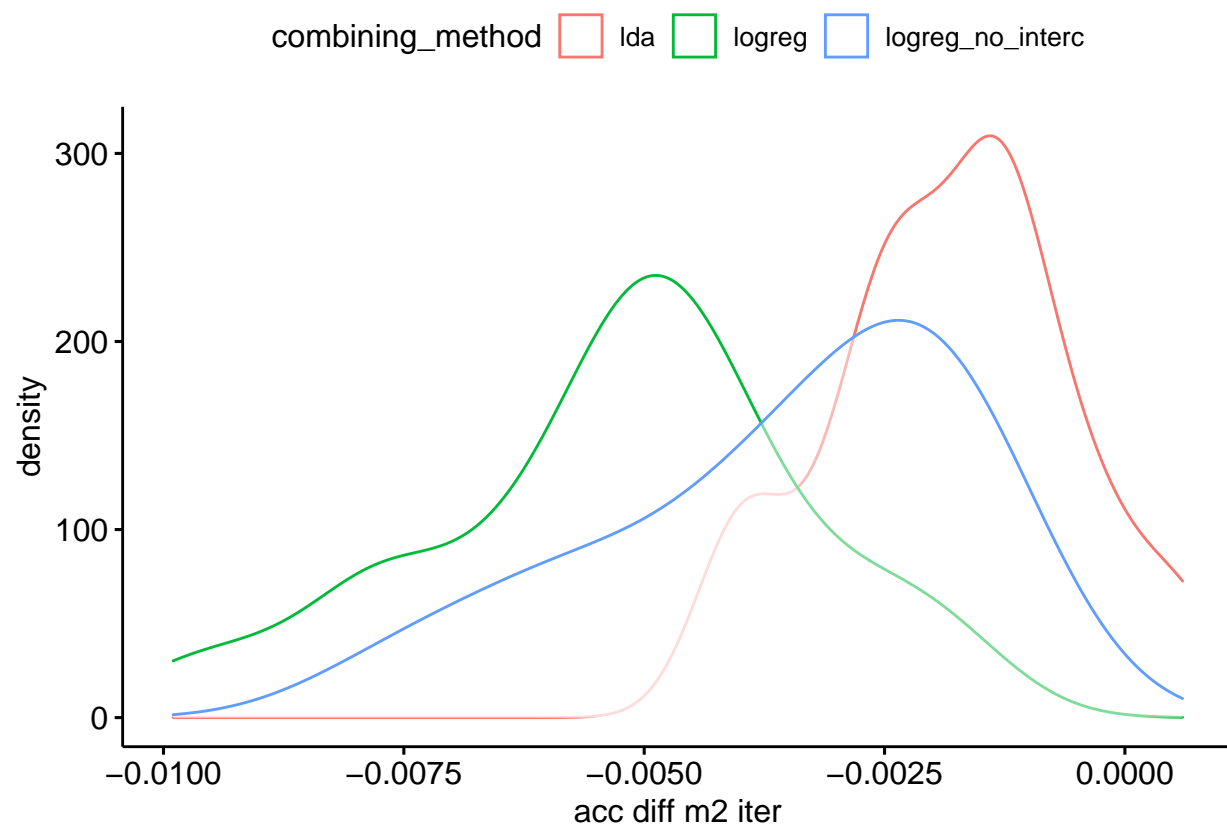


```
ggqqplot(differences, x="acc_diff_m2", color="combining_method")
```

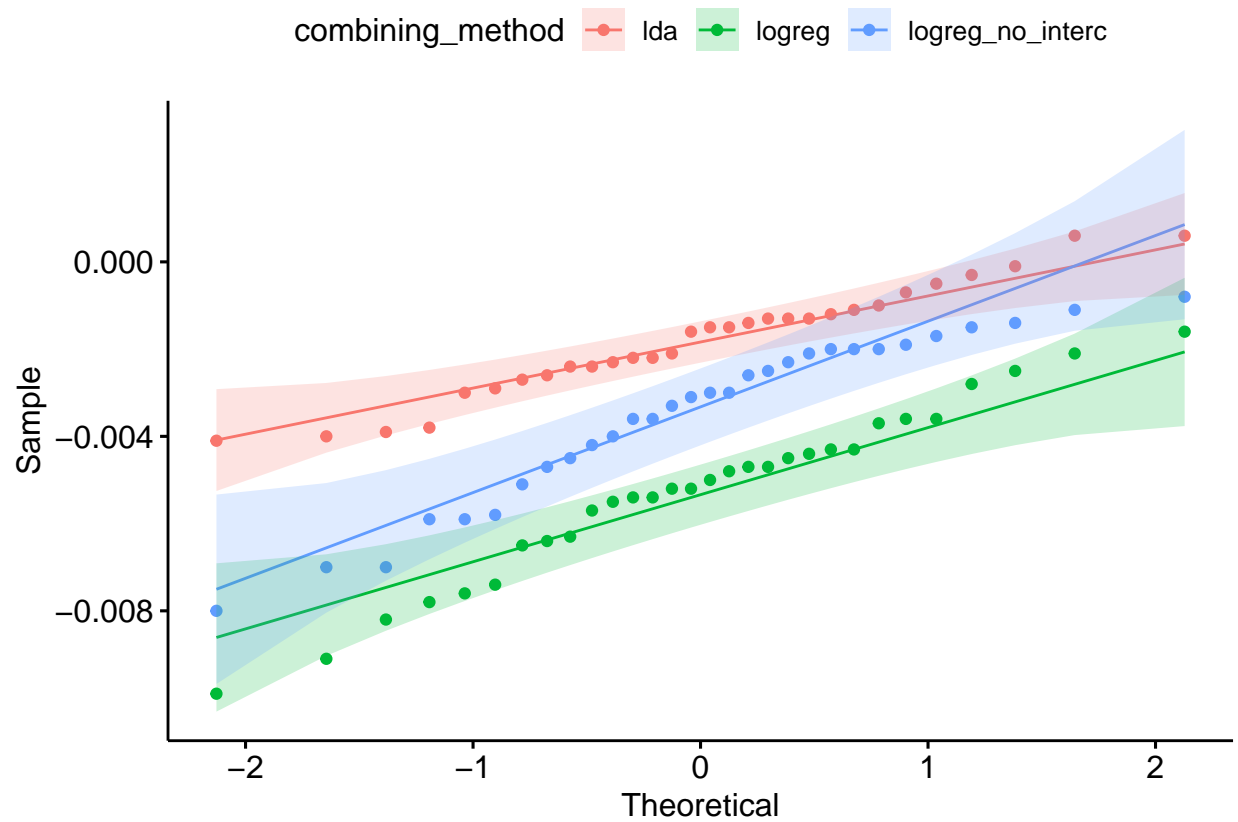




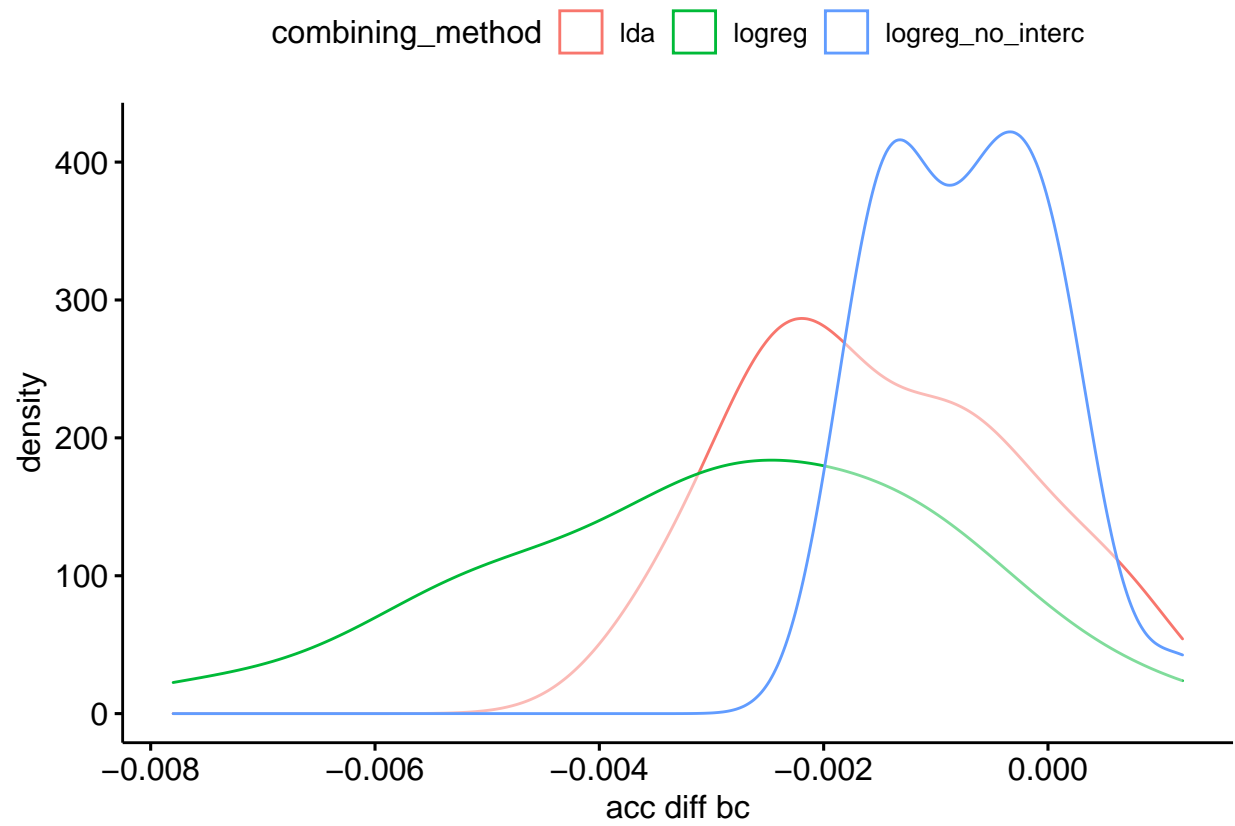
```
ggdensity(data=differences, x="acc_diff_m2_iter", color="combining_method", xlab="acc diff m2 iter")
```



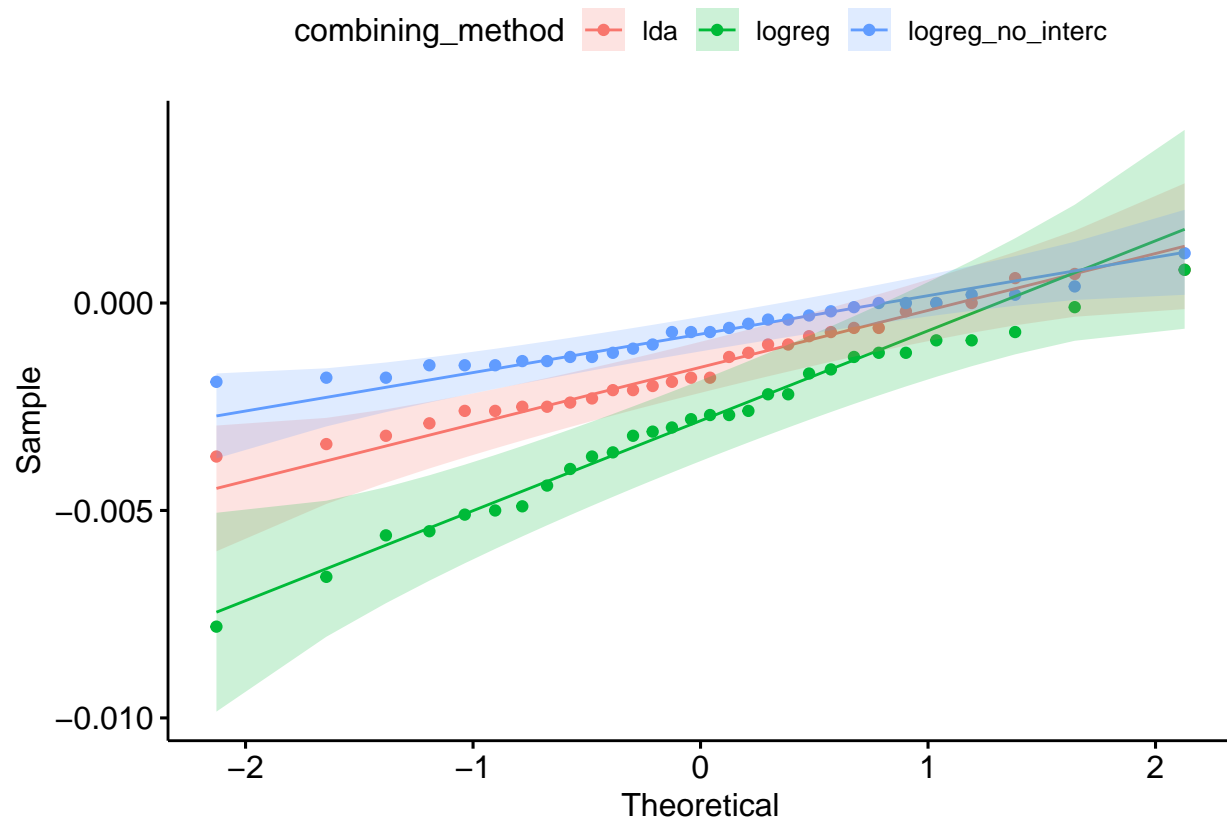
```
ggqqplot(differences, x="acc_diff_m2_iter", color="combining_method")
```



```
ggdensity(data=differences, x="acc_diff_bc", color="combining_method", xlab="acc diff bc")
```



```
ggqqplot(differences, x="acc_diff_bc", color="combining_method")
```



```
lda_diff <- differences %>% filter(combining_method=="lda")
shapiro.test(lda_diff$acc_diff_bc)
```

```
##
## Shapiro-Wilk normality test
##
## data:  lda_diff$acc_diff_bc
## W = 0.96681, p-value = 0.456
```

```
logreg_diff <- differences %>% filter(combining_method=="logreg")
shapiro.test(logreg_diff$acc_diff_bc)
```

```
##
## Shapiro-Wilk normality test
##
## data:  logreg_diff$acc_diff_bc
## W = 0.98069, p-value = 0.8436
```

```
logreg_ni_diff <- differences %>% filter(combining_method=="logreg_no_interc")
shapiro.test(logreg_ni_diff$acc_diff_bc)
```

```
##
## Shapiro-Wilk normality test
##
```

```
## data: logreg_ni_diff$acc_diff_bc
## W = 0.96103, p-value = 0.329
```

Shapiro-Wilk zero hypothesis is data normality. Accuracy differences seem to be normal, so we can proceed with the t-tests.

## LDA

```
ens_results_wide_lda <- ens_results_wide %>% filter(combining_method=="lda")
t.test(ens_results_wide_lda$accuracy_vt_m1, ens_results_wide_lda$accuracy_tt_m1, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_m1 and ens_results_wide_lda$accuracy_tt_m1
## t = -7.8715, df = 29, p-value = 1.11e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.002242492 -0.001317508
## sample estimates:
## mean of the differences
## -0.00178
```

```
t.test(ens_results_wide_lda$accuracy_vt_m1, ens_results_wide_lda$accuracy_tt_m1, paired=TRUE, alternative="two.sided")
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_m1 and ens_results_wide_lda$accuracy_tt_m1
## t = -7.8715, df = 29, p-value = 5.551e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -0.001395773
## sample estimates:
## mean of the differences
## -0.00178
```

```
t.test(ens_results_wide_lda$accuracy_vt_m2, ens_results_wide_lda$accuracy_tt_m2, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_m2 and ens_results_wide_lda$accuracy_tt_m2
## t = -7.8622, df = 29, p-value = 1.137e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.002272440 -0.001334227
## sample estimates:
## mean of the differences
## -0.001803333
```

```
t.test(ens_results_wide_lda$accuracy_vt_m2, ens_results_wide_lda$accuracy_tt_m2, paired=TRUE, alternati
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_m2 and ens_results_wide_lda$accuracy_tt_m2
## t = -7.8622, df = 29, p-value = 5.684e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -0.001413611
## sample estimates:
## mean of the differences
##      -0.001803333
```

```
t.test(ens_results_wide_lda$accuracy_vt_m2_iter, ens_results_wide_lda$accuracy_tt_m2_iter, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_m2_iter and ens_results_wide_lda$accuracy_tt_m2_iter
## t = -7.8401, df = 29, p-value = 1.203e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.002277966 -0.001335367
## sample estimates:
## mean of the differences
##      -0.001806667
```

```
t.test(ens_results_wide_lda$accuracy_vt_m2_iter, ens_results_wide_lda$accuracy_tt_m2_iter, paired=TRUE,
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_m2_iter and ens_results_wide_lda$accuracy_tt_m2_iter
## t = -7.8401, df = 29, p-value = 6.013e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -0.001415122
## sample estimates:
## mean of the differences
##      -0.001806667
```

```
t.test(ens_results_wide_lda$accuracy_vt_bc, ens_results_wide_lda$accuracy_tt_bc, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_bc and ens_results_wide_lda$accuracy_tt_bc
## t = -6.6405, df = 29, p-value = 2.794e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
## -0.001966353 -0.001040314
## sample estimates:
## mean of the differences
## -0.001503333
```

```
t.test(ens_results_wide_lda$accuracy_vt_bc, ens_results_wide_lda$accuracy_tt_bc, paired=TRUE, alternati
```

```
##
## Paired t-test
##
## data: ens_results_wide_lda$accuracy_vt_bc and ens_results_wide_lda$accuracy_tt_bc
## t = -6.6405, df = 29, p-value = 1.397e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -0.001118668
## sample estimates:
## mean of the differences
## -0.001503333
```

## LogReg

```
ens_results_wide_logreg <- ens_results_wide %>% filter(combining_method=="logreg")
t.test(ens_results_wide_logreg$accuracy_vt_m1, ens_results_wide_logreg$accuracy_tt_m1, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_logreg$accuracy_vt_m1 and ens_results_wide_logreg$accuracy_tt_m1
## t = -13.458, df = 29, p-value = 5.324e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.005533312 -0.004073354
## sample estimates:
## mean of the differences
## -0.004803333
```

```
t.test(ens_results_wide_logreg$accuracy_vt_m1, ens_results_wide_logreg$accuracy_tt_m1, paired=TRUE, alt
```

```
##
## Paired t-test
##
## data: ens_results_wide_logreg$accuracy_vt_m1 and ens_results_wide_logreg$accuracy_tt_m1
## t = -13.458, df = 29, p-value = 2.662e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -0.004196884
## sample estimates:
## mean of the differences
## -0.004803333
```



```
t.test(ens_results_wide_logreg$accuracy_vt_m2, ens_results_wide_logreg$accuracy_tt_m2, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_logreg$accuracy_vt_m2 and ens_results_wide_logreg$accuracy_tt_m2
## t = -14.523, df = 29, p-value = 7.704e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.006015945 -0.004530722
## sample estimates:
## mean of the differences
## -0.005273333
```

```
t.test(ens_results_wide_logreg$accuracy_vt_m2, ens_results_wide_logreg$accuracy_tt_m2, paired=TRUE, alt="less")
```

```
##
## Paired t-test
##
## data: ens_results_wide_logreg$accuracy_vt_m2 and ens_results_wide_logreg$accuracy_tt_m2
## t = -14.523, df = 29, p-value = 3.852e-15
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -0.00465639
## sample estimates:
## mean of the differences
## -0.005273333
```

```
t.test(ens_results_wide_logreg$accuracy_vt_m2_iter, ens_results_wide_logreg$accuracy_tt_m2_iter, paired=TRUE)
```

```
##
## Paired t-test
##
## data: ens_results_wide_logreg$accuracy_vt_m2_iter and ens_results_wide_logreg$accuracy_tt_m2_iter
## t = -14.523, df = 29, p-value = 7.704e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.006015945 -0.004530722
## sample estimates:
## mean of the differences
## -0.005273333
```

```
t.test(ens_results_wide_logreg$accuracy_vt_m2_iter, ens_results_wide_logreg$accuracy_tt_m2_iter, paired=TRUE, alt="less")
```

```
##
## Paired t-test
##
## data: ens_results_wide_logreg$accuracy_vt_m2_iter and ens_results_wide_logreg$accuracy_tt_m2_iter
## t = -14.523, df = 29, p-value = 3.852e-15
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
```

```
##          -Inf -0.00465639
## sample estimates:
## mean of the differences
##          -0.005273333

t.test(ens_results_wide_logreg$accuracy_vt_bc, ens_results_wide_logreg$accuracy_tt_bc, paired=TRUE)

##
## Paired t-test
##
## data:  ens_results_wide_logreg$accuracy_vt_bc and ens_results_wide_logreg$accuracy_tt_bc
## t = -8.1519, df = 29, p-value = 5.464e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.003731823 -0.002234844
## sample estimates:
## mean of the differences
##          -0.002983333

t.test(ens_results_wide_logreg$accuracy_vt_bc, ens_results_wide_logreg$accuracy_tt_bc, paired=TRUE, alt="not.equal")

##
## Paired t-test
##
## data:  ens_results_wide_logreg$accuracy_vt_bc and ens_results_wide_logreg$accuracy_tt_bc
## t = -8.1519, df = 29, p-value = 2.732e-09
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.002361506
## sample estimates:
## mean of the differences
##          -0.002983333
```

### LogReg without intercept

```
ens_results_wide_logreg_ni <- ens_results_wide %>% filter(combining_method=="logreg_no_interc")
t.test(ens_results_wide_logreg_ni$accuracy_vt_m1, ens_results_wide_logreg_ni$accuracy_tt_m1, paired=TRUE)

##
## Paired t-test
##
## data:  ens_results_wide_logreg_ni$accuracy_vt_m1 and ens_results_wide_logreg_ni$accuracy_tt_m1
## t = -9.0337, df = 29, p-value = 6.281e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.003499328 -0.002207339
## sample estimates:
## mean of the differences
##          -0.002853333
```

```

t.test(ens_results_wide_logreg_ni$accuracy_vt_m1, ens_results_wide_logreg_ni$accuracy_tt_m1, paired=TRUE)

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_m1 and ens_results_wide_logreg_ni$accuracy_tt_m1
## t = -9.0337, df = 29, p-value = 3.141e-10
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -0.002316657
## sample estimates:
## mean of the differences
##      -0.002853333

t.test(ens_results_wide_logreg_ni$accuracy_vt_m2, ens_results_wide_logreg_ni$accuracy_tt_m2, paired=TRUE)

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_m2 and ens_results_wide_logreg_ni$accuracy_tt_m2
## t = -10.026, df = 29, p-value = 6.215e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##      -0.00423802 -0.00280198
## sample estimates:
## mean of the differences
##      -0.00352

t.test(ens_results_wide_logreg_ni$accuracy_vt_m2, ens_results_wide_logreg_ni$accuracy_tt_m2, paired=TRUE)

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_m2 and ens_results_wide_logreg_ni$accuracy_tt_m2
## t = -10.026, df = 29, p-value = 3.108e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -0.002923486
## sample estimates:
## mean of the differences
##      -0.00352

t.test(ens_results_wide_logreg_ni$accuracy_vt_m2_iter, ens_results_wide_logreg_ni$accuracy_tt_m2_iter, paired=TRUE)

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_m2_iter and ens_results_wide_logreg_ni$accuracy_tt_m2_iter
## t = -10.026, df = 29, p-value = 6.215e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:

```

```

## -0.00423802 -0.00280198
## sample estimates:
## mean of the differences
## -0.00352

t.test(ens_results_wide_logreg_ni$accuracy_vt_m2_iter, ens_results_wide_logreg_ni$accuracy_tt_m2_iter,

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_m2_iter and ens_results_wide_logreg_ni$accuracy_tt_m2_
## t = -10.026, df = 29, p-value = 3.108e-11
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -0.002923486
## sample estimates:
## mean of the differences
## -0.00352

t.test(ens_results_wide_logreg_ni$accuracy_vt_bc, ens_results_wide_logreg_ni$accuracy_tt_bc, paired=TRUE

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_bc and ens_results_wide_logreg_ni$accuracy_tt_bc
## t = -5.0961, df = 29, p-value = 1.942e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.0009949435 -0.0004250565
## sample estimates:
## mean of the differences
## -0.00071

t.test(ens_results_wide_logreg_ni$accuracy_vt_bc, ens_results_wide_logreg_ni$accuracy_tt_bc, paired=TRUE

##
## Paired t-test
##
## data: ens_results_wide_logreg_ni$accuracy_vt_bc and ens_results_wide_logreg_ni$accuracy_tt_bc
## t = -5.0961, df = 29, p-value = 9.709e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
## -Inf -0.0004732759
## sample estimates:
## mean of the differences
## -0.00071

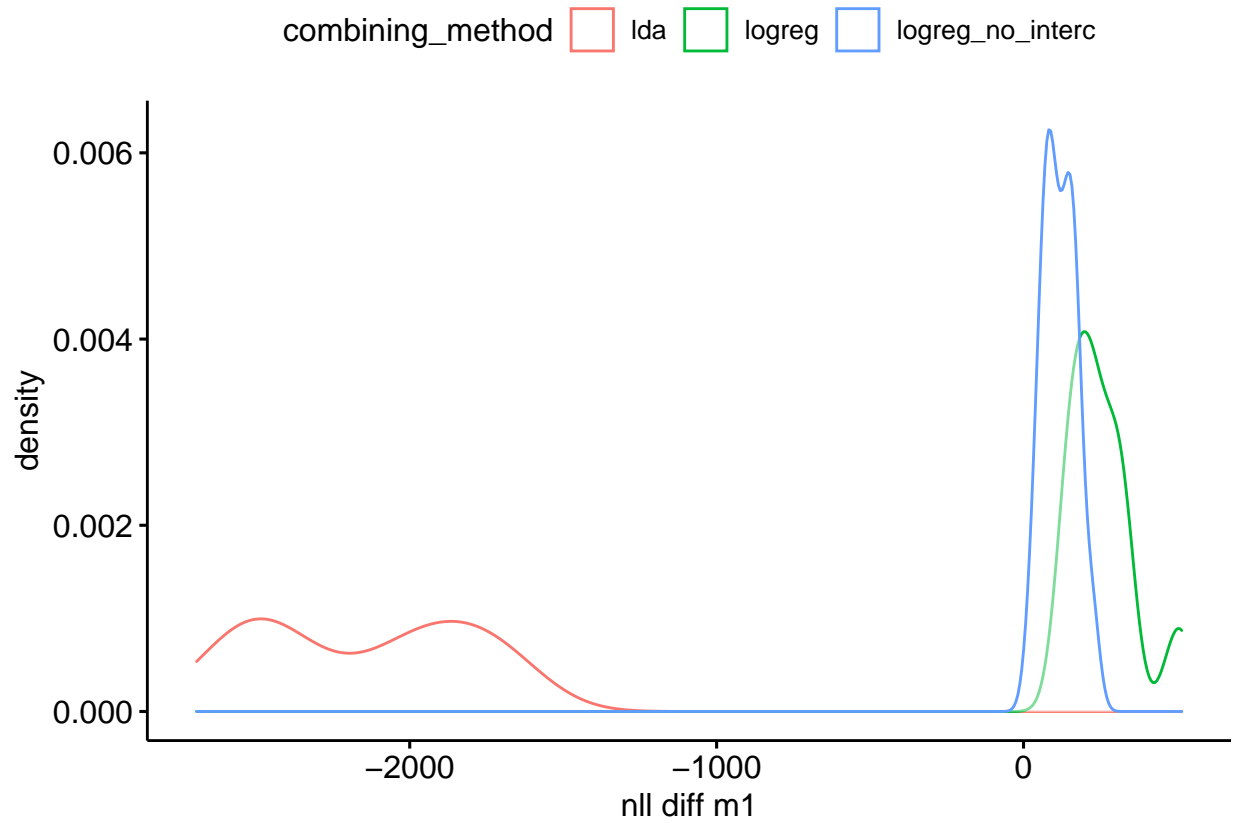
```

For all combining methods and for all coupling methods, we found, that ensemble models trained on subset (of size 500) of the neural networks training set have statistically significantly better accuracy than ensemble models trained on separate validation set (of the same size).

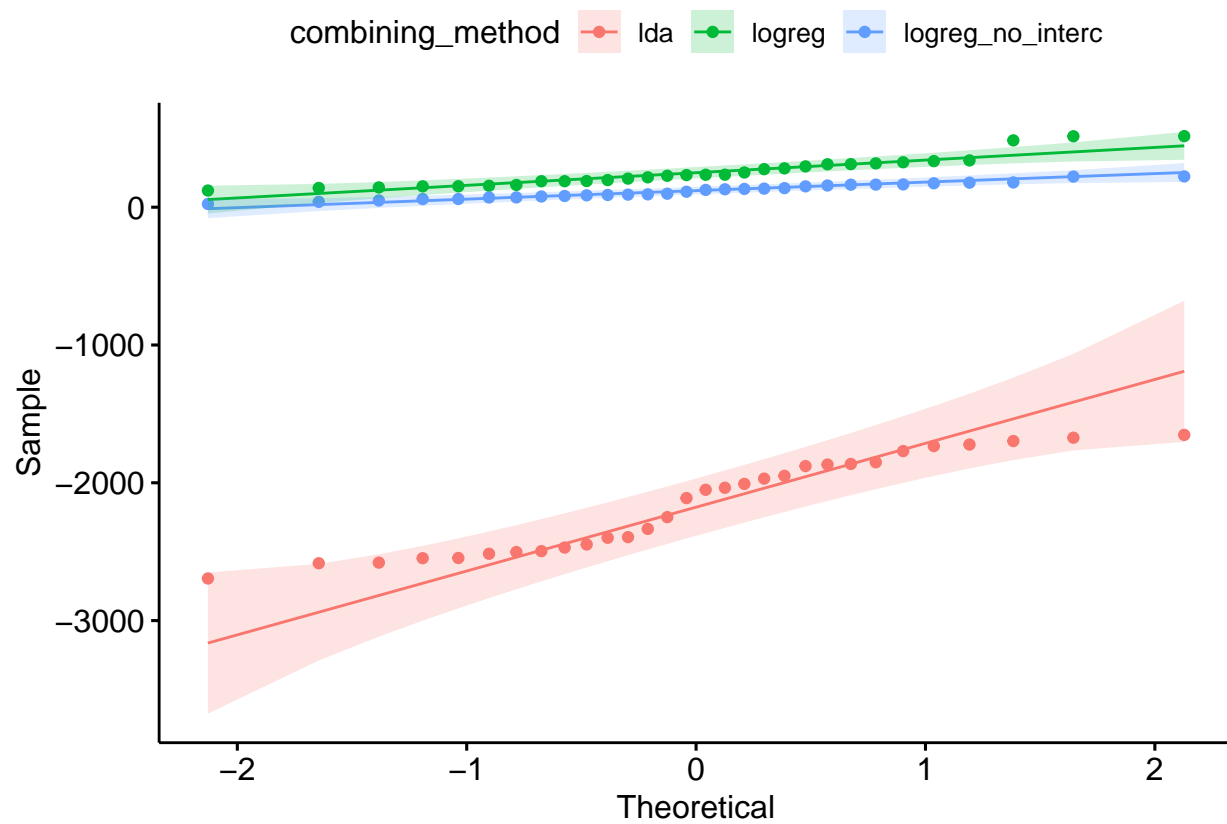
## Negative log likelihood

Ascertaining normality of differences

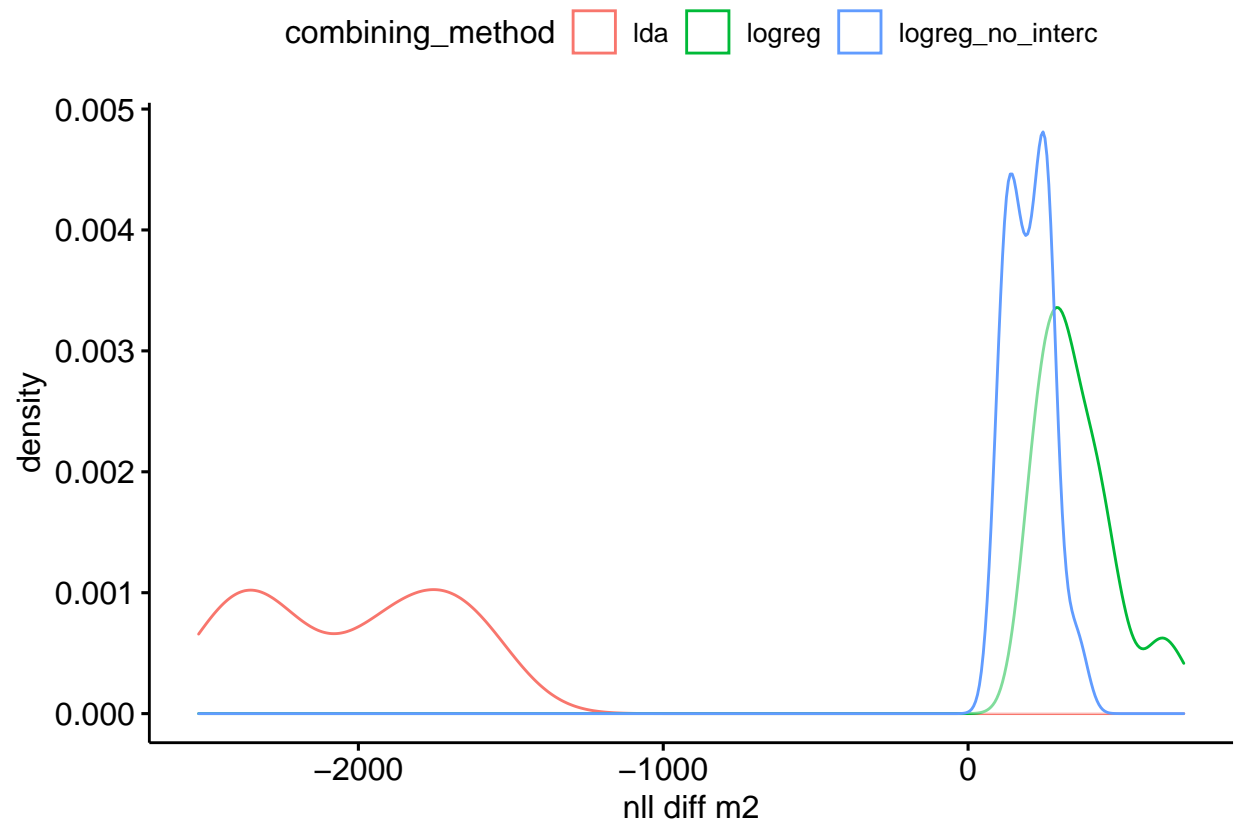
```
ggdensity(data=differences, x="nll_diff_m1", color="combining_method", xlab="nll diff m1")
```



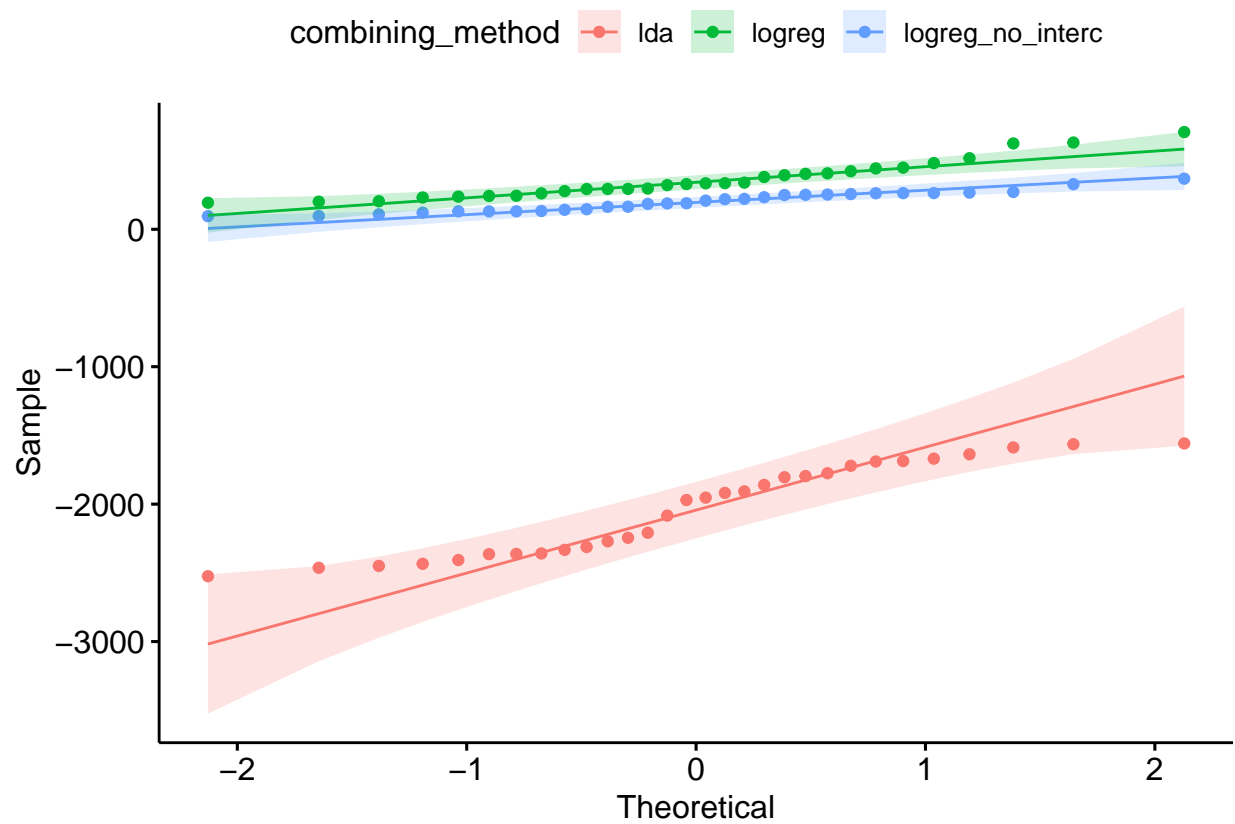
```
ggqqplot(differences, x="nll_diff_m1", color="combining_method")
```



```
ggdensity(data=differences, x="nll_diff_m2", color="combining_method", xlab="nll diff m2")
```

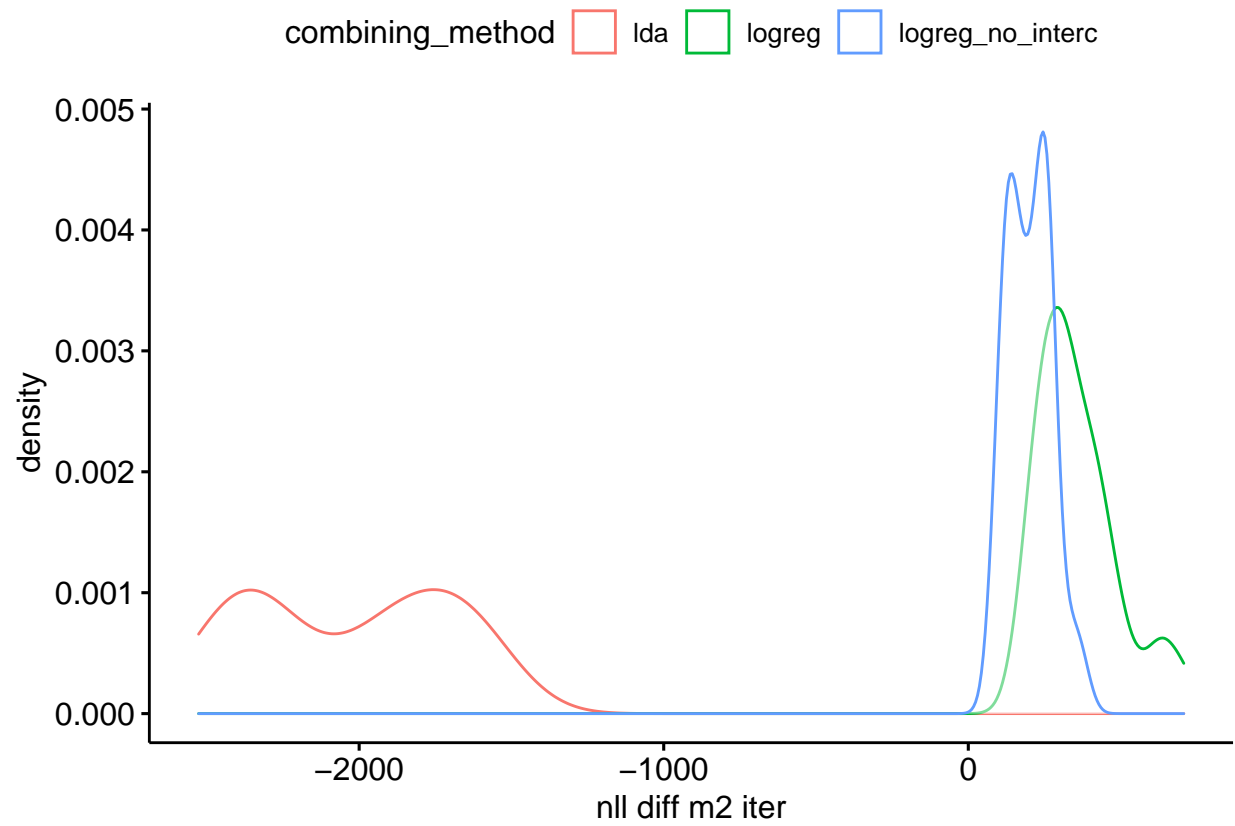


```
ggqqplot(differences, x="nll_diff_m2", color="combining_method")
```

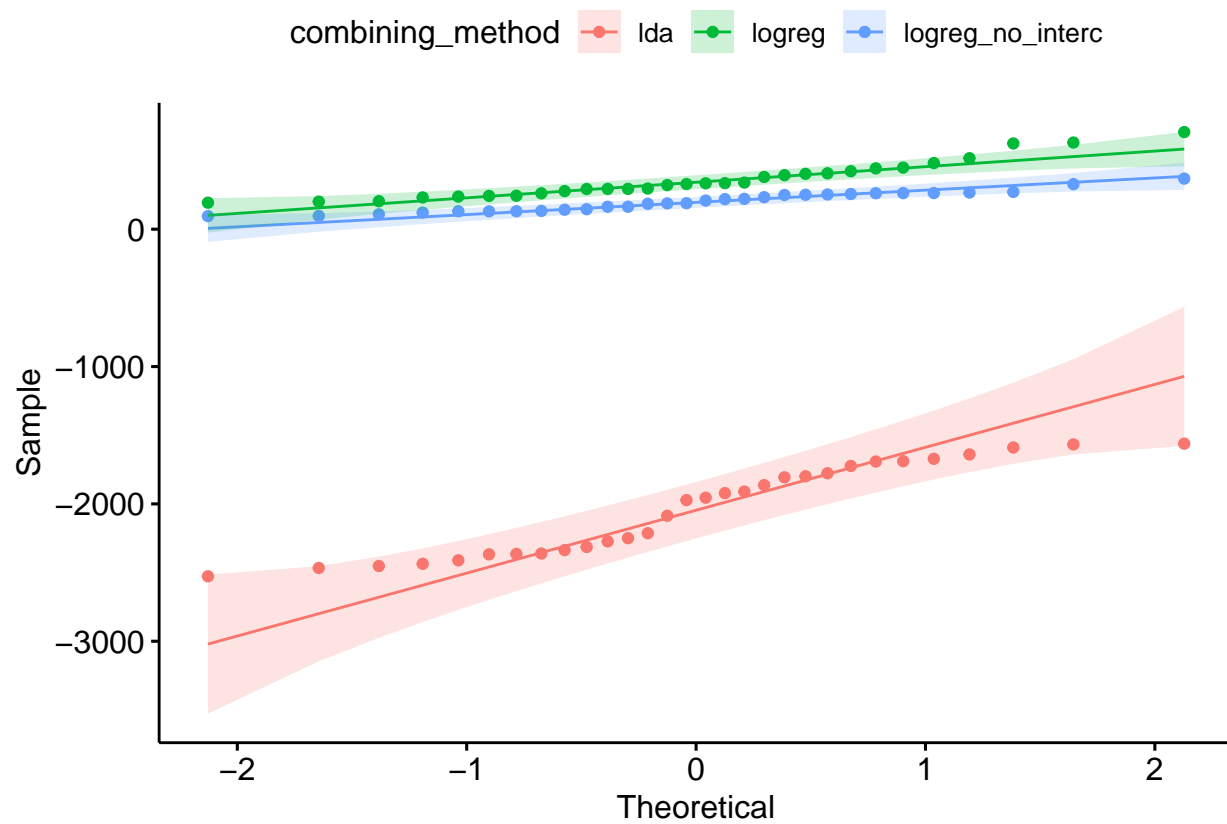


```
ggdensity(data=differences, x="nll_diff_m2_iter", color="combining_method", xlab="nll diff m2 iter")
```

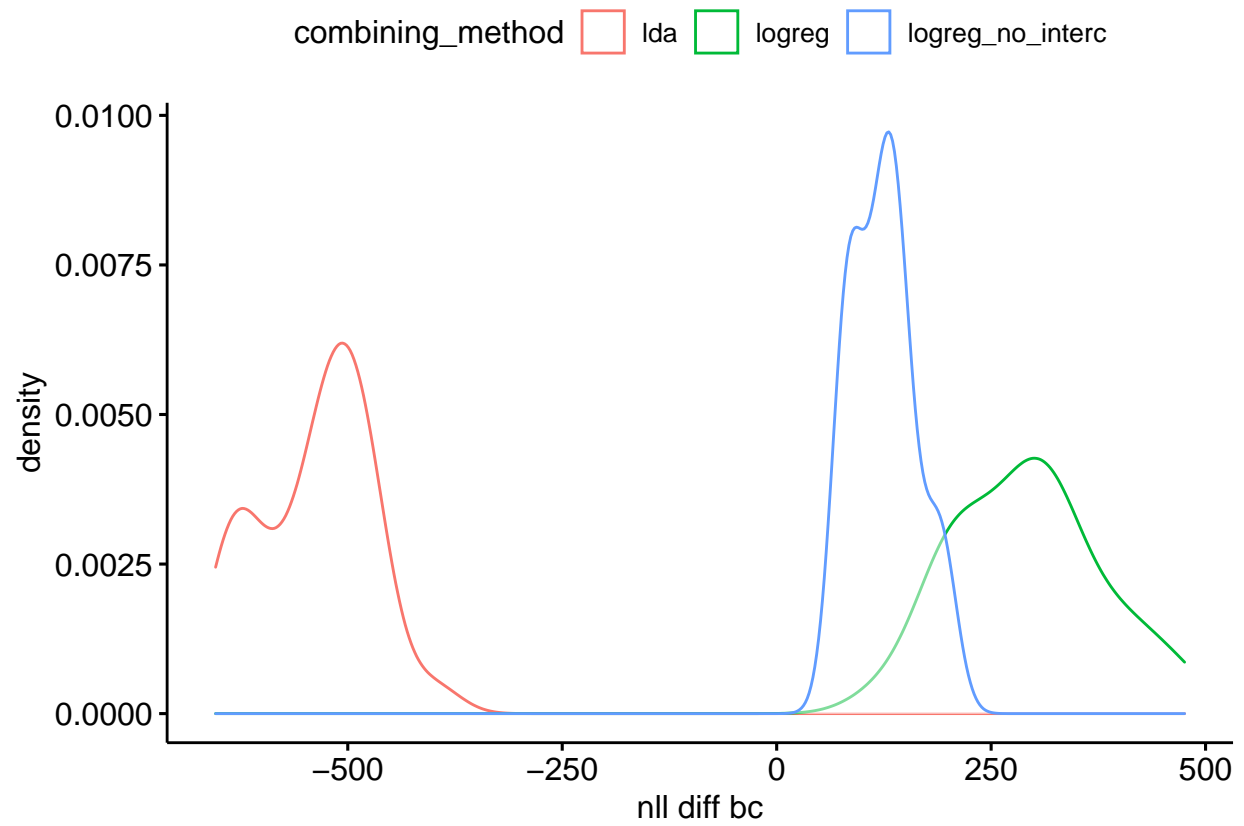




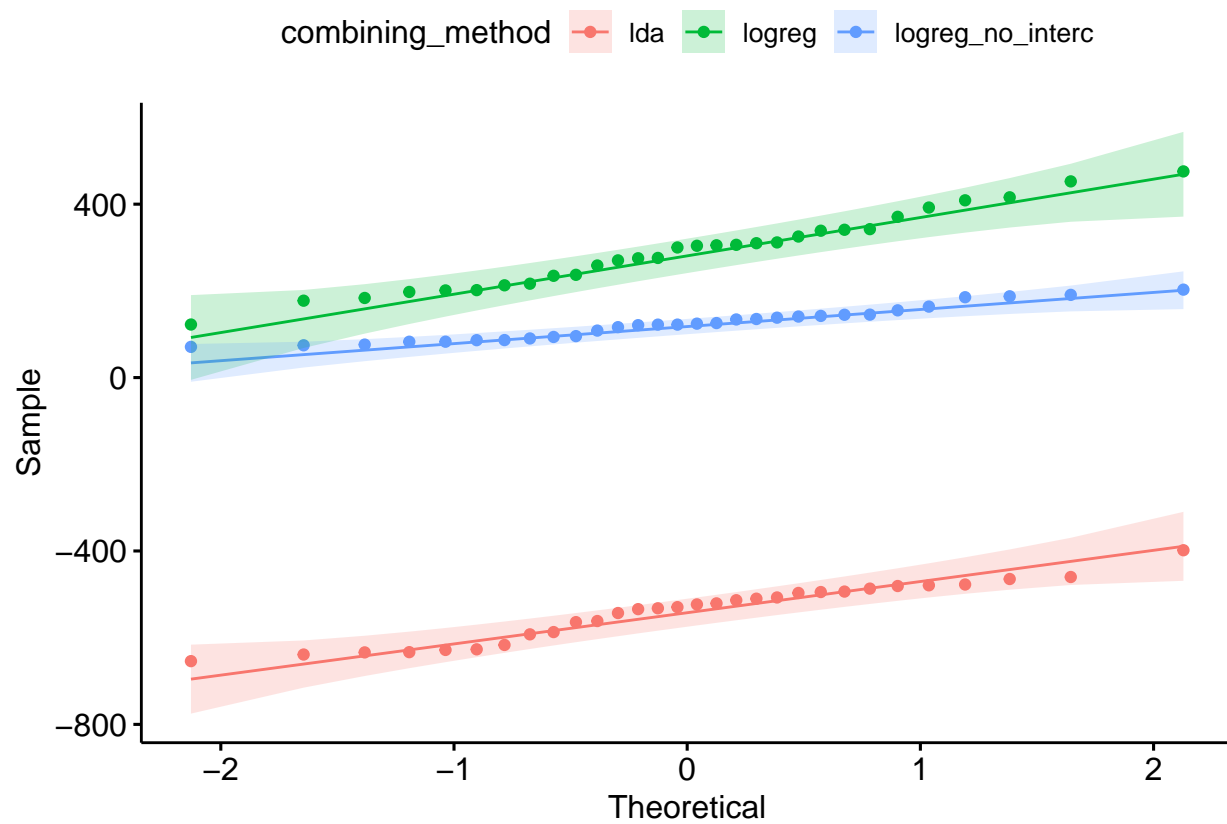
```
ggqqplot(differences, x="nll_diff_m2_iter", color="combining_method")
```



```
ggdensity(data=differences, x="nll_diff_bc", color="combining_method", xlab="nll diff bc")
```



```
ggqqplot(differences, x="nll_diff_bc", color="combining_method")
```



Differences don't seem to be normal in several cases. We postpone tests on this secondary metric.