LDA training on random subsets of different sizes

```
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'
```

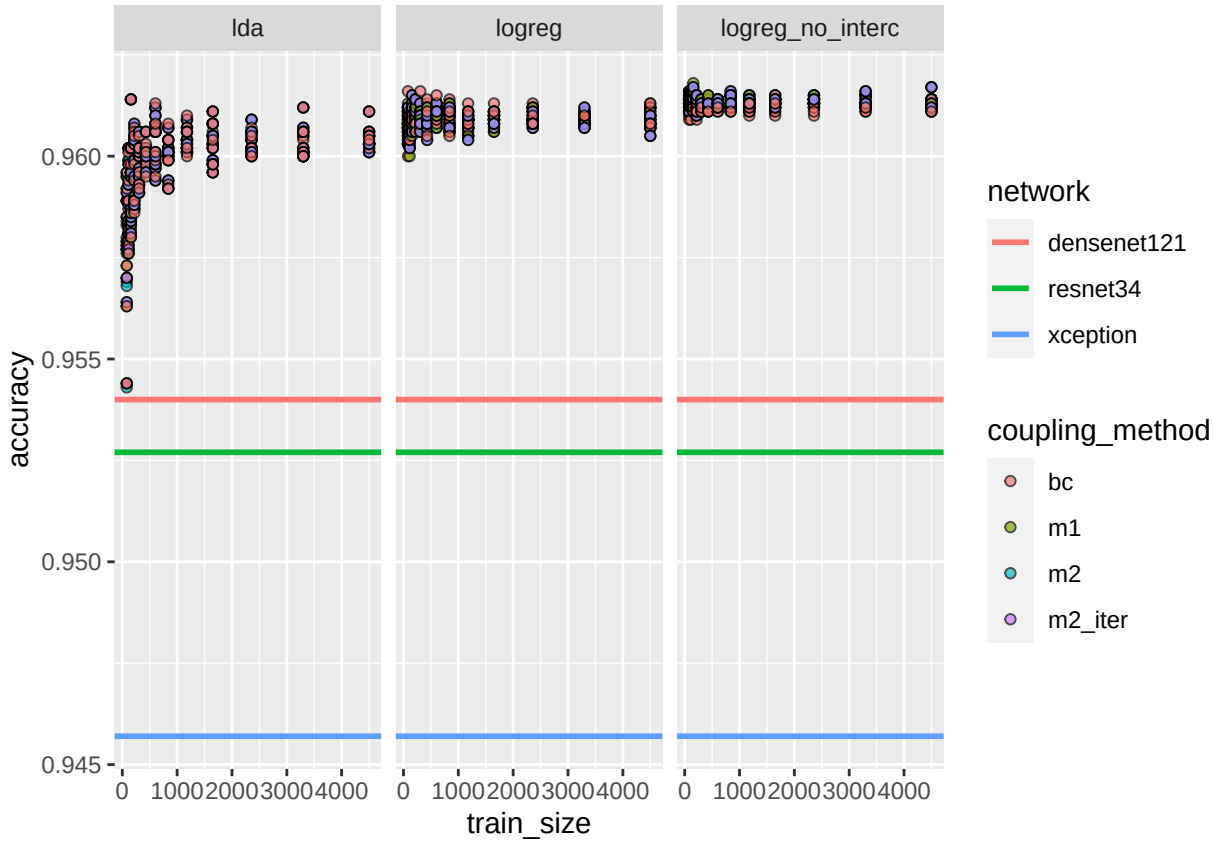
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

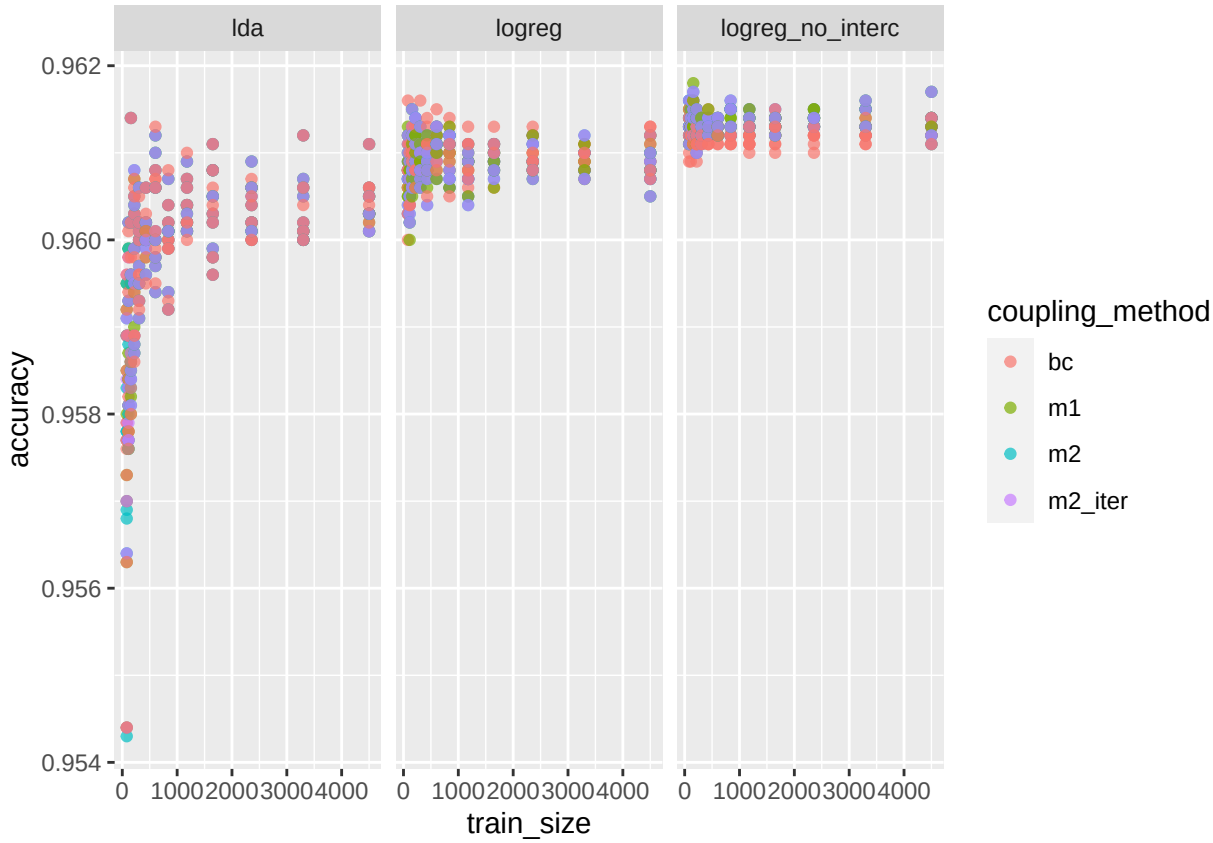
Experiment code is in the file training_subsets_sizes_experiment.py. Experiment on CIFAR10 dataset. This experiment trains WeightedLinearEnsemble on various subsets of different sizes of data on which neural networks were trained. Subsets of the same size are disjoint. Three different coupling methods are used: method one and two from (Wu, Lin, and Weng 2004) and Bayes covariant method (Šuch and Barreda 2016). Goal of this experiment is to determine, for which size of the LDA training set, the ensemble achieves the best performance.

```
acc_ens_subsets <- read.csv("../data/data_train_val_c10/0/exp_subsets_sizes_train_outputs/accuracies.csv
acc_nets <- read.csv("../data/data_train_val_c10/0/exp_subsets_sizes_train_outputs/net_accuracies.csv",
```

```
scatter <- ggplot() + geom_point(data=acc_ens_subsets, mapping=aes(x=train_size, y=accuracy, fill=coupli
scatter
```



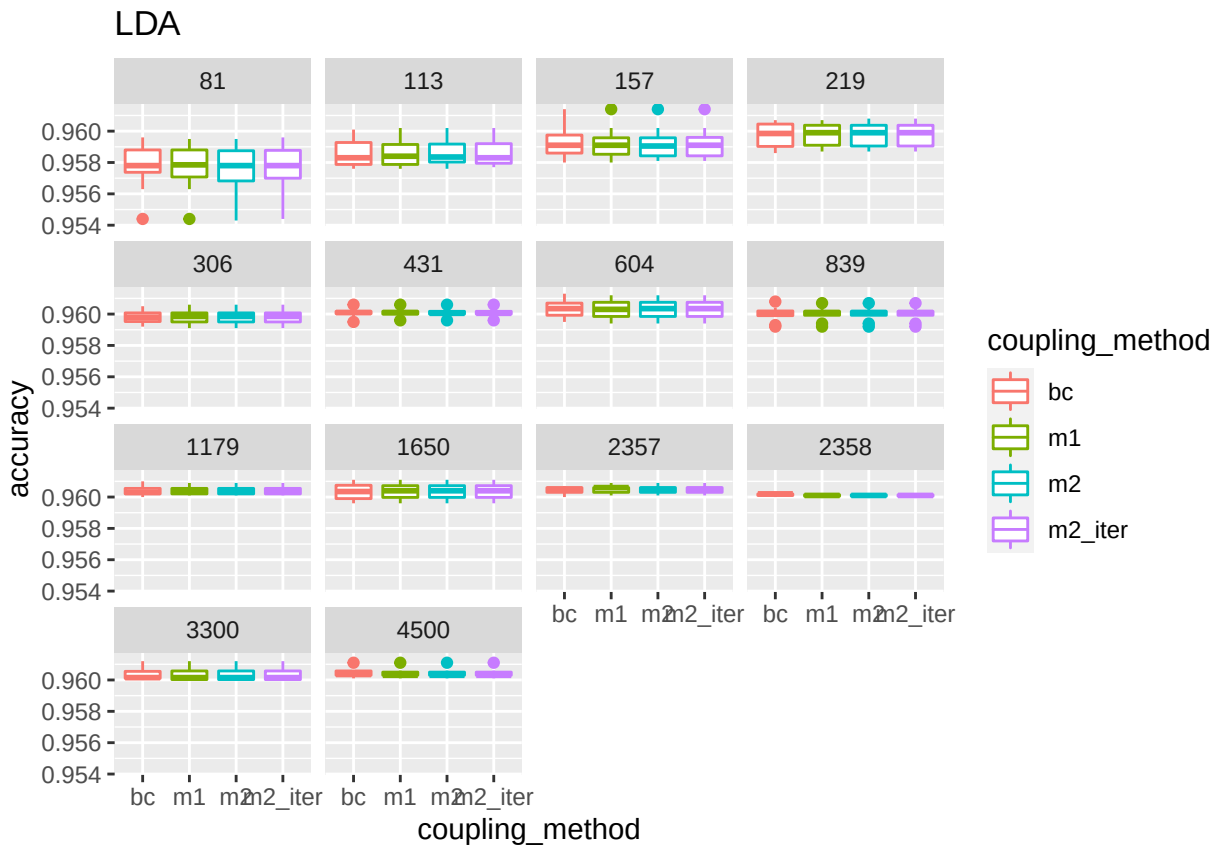

```
scatter2 <- ggplot() + geom_point(data=acc_ens_subsets, mapping=aes(x=train_size, y=accuracy, color=cou
scatter2
```



Scatter plots are quite messy, so we will use boxplots.

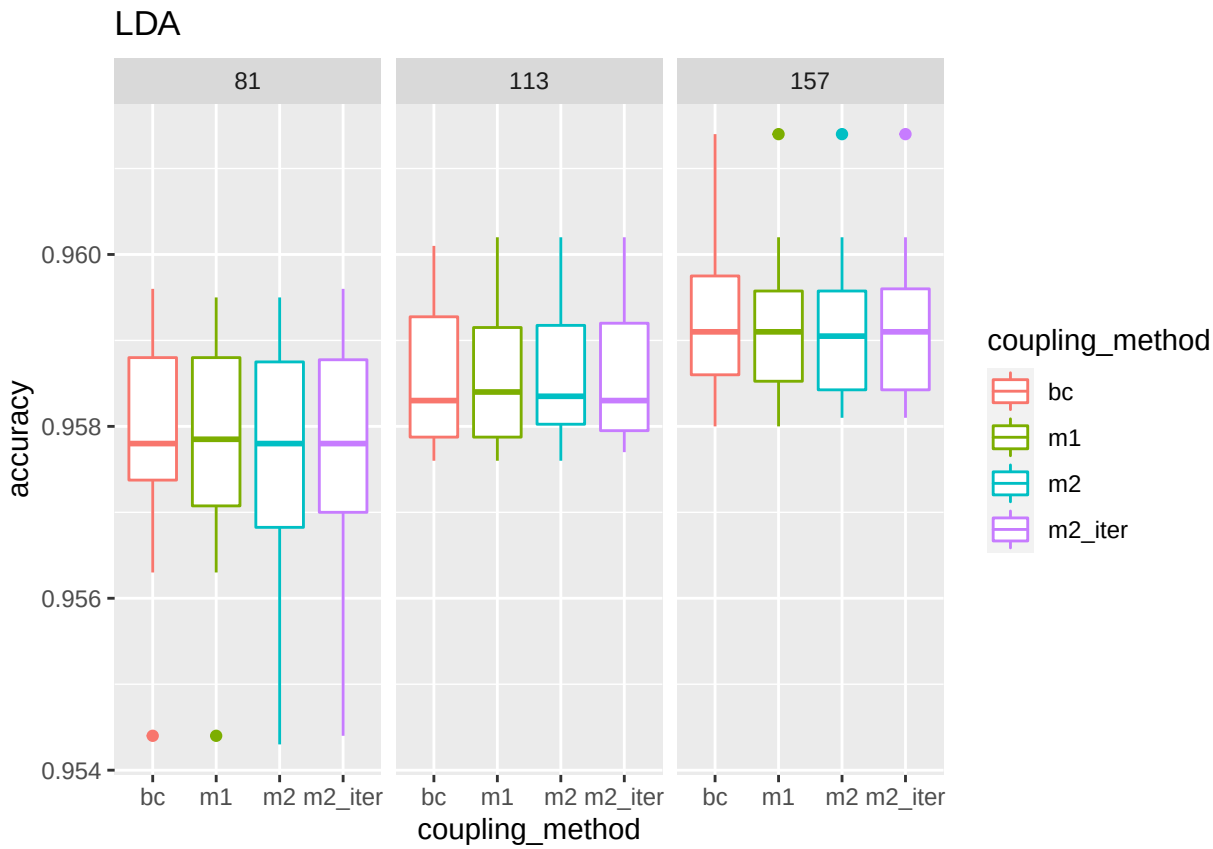
LDA

```
box_acc <- acc_ens_subsets %>% filter(combining_method=="lda") %>% ggplot() + geom_boxplot(mapping=aes(
box_acc
```

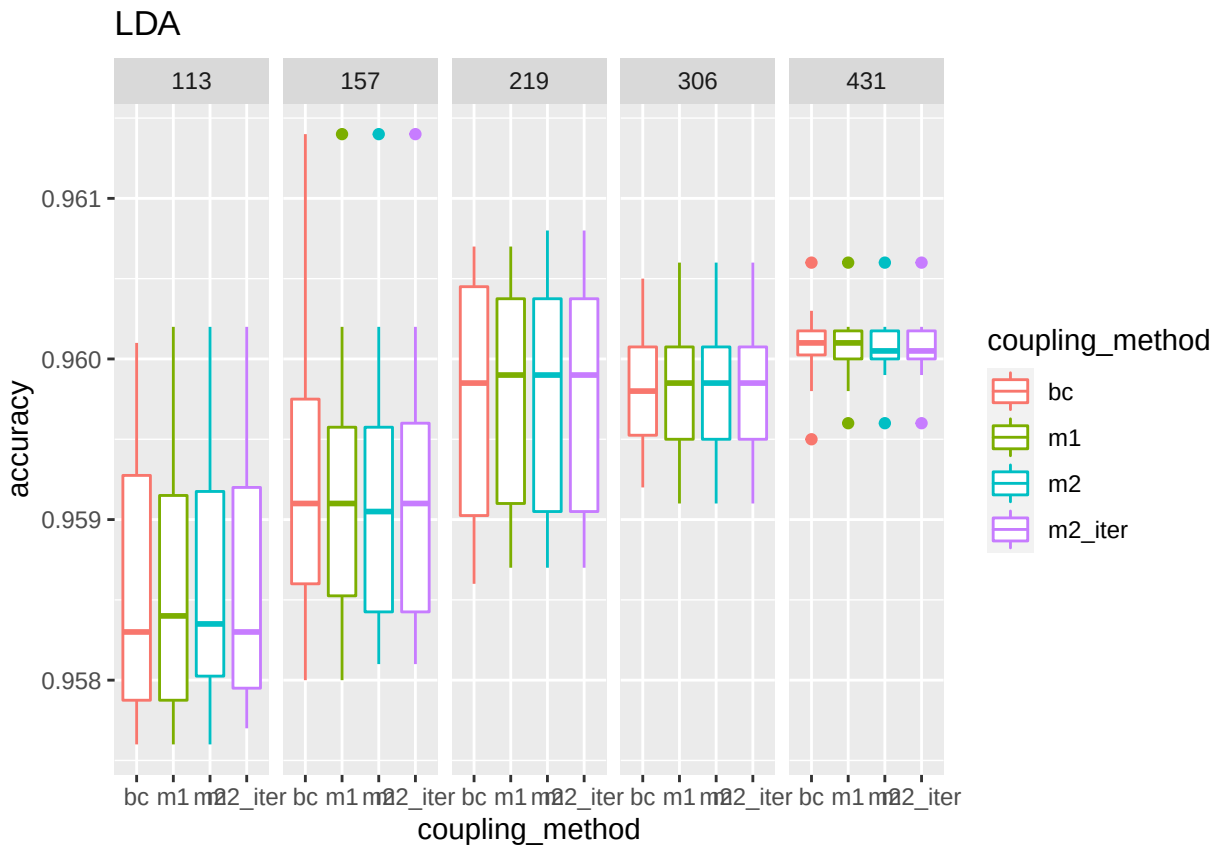


Accuracy seems to be slowly increasing with increasing training set size. We will inspect this closer by plotting for smaller subset sizes range.

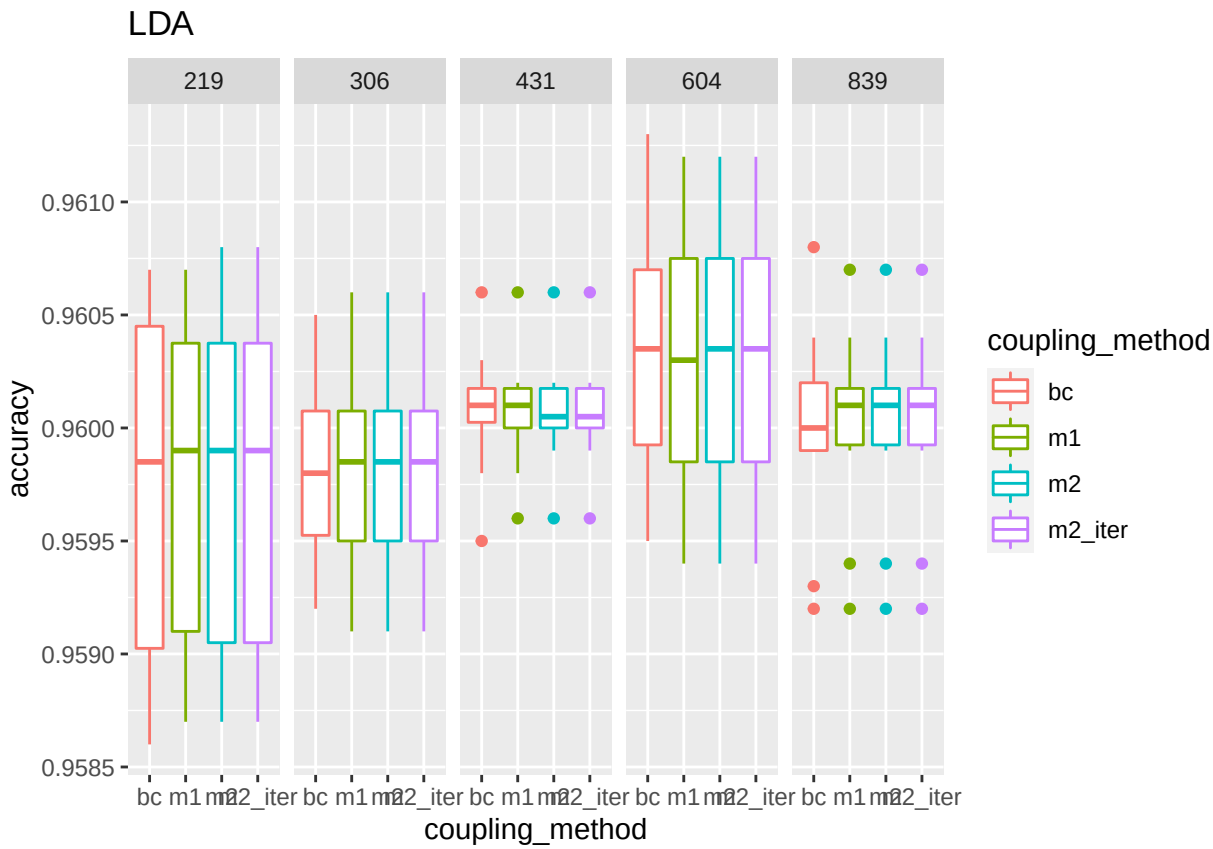
```
acc_ens_subsets %>% filter(0 < train_size & train_size < 200 & combining_method=="lda") %>% ggplot() + g
```




```
acc_ens_subsets %>% filter(100 < train_size & train_size < 500 & combining_method=="lda") %>% ggplot() +
```

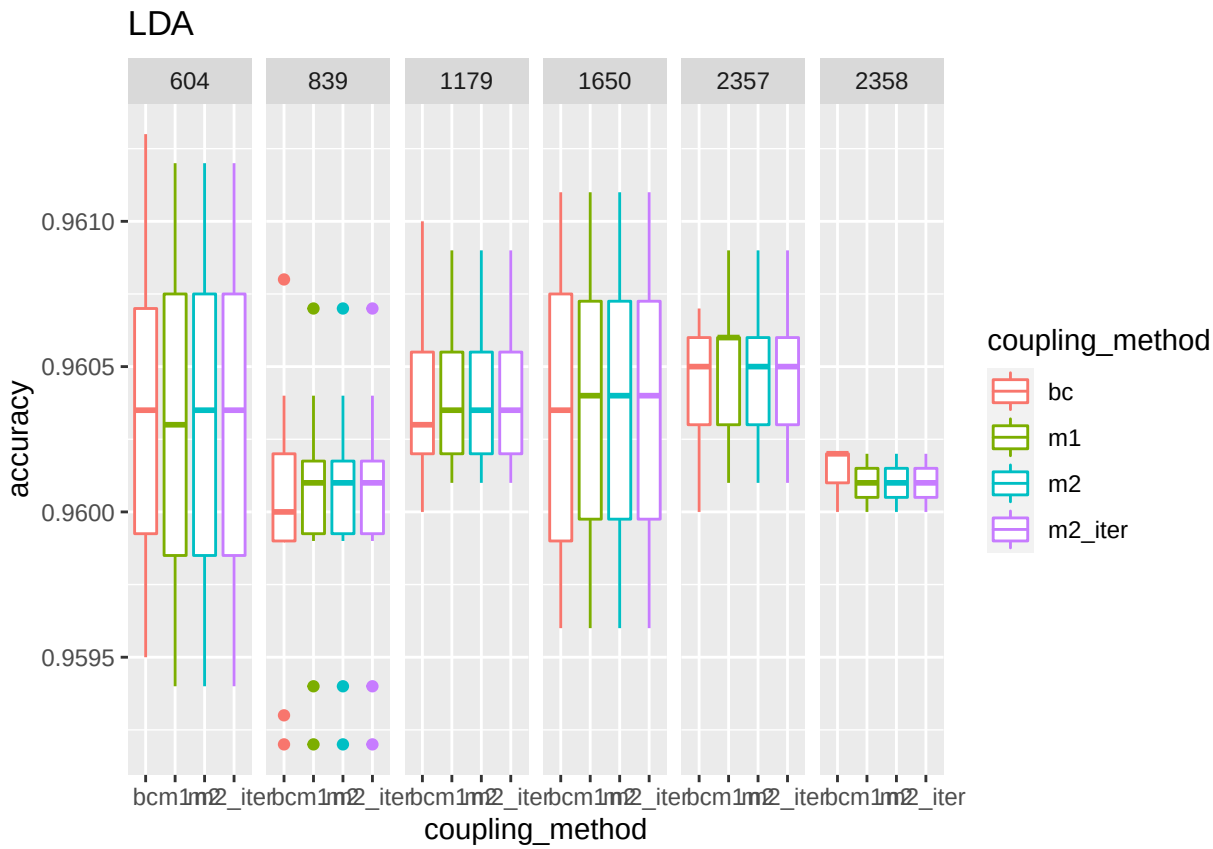



```
acc_ens_subsets %>% filter(200 < train_size & train_size < 1000 & combining_method=="lda") %>% ggplot()
```



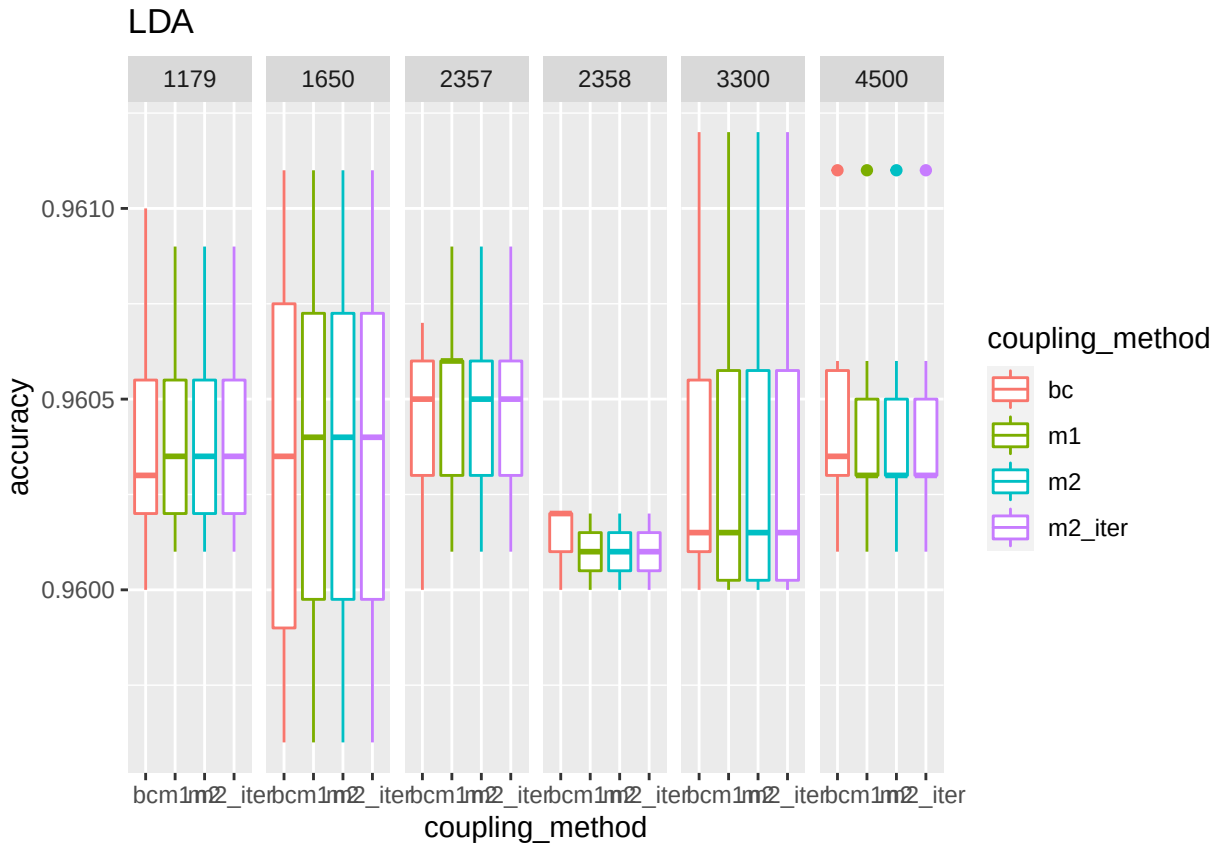
Here the accuracy increase seems to be stopping at train set size 604.

```
acc_ens_subsets %>% filter(500 < train_size & train_size < 3000 & combining_method=="lda") %>% ggplot()
```



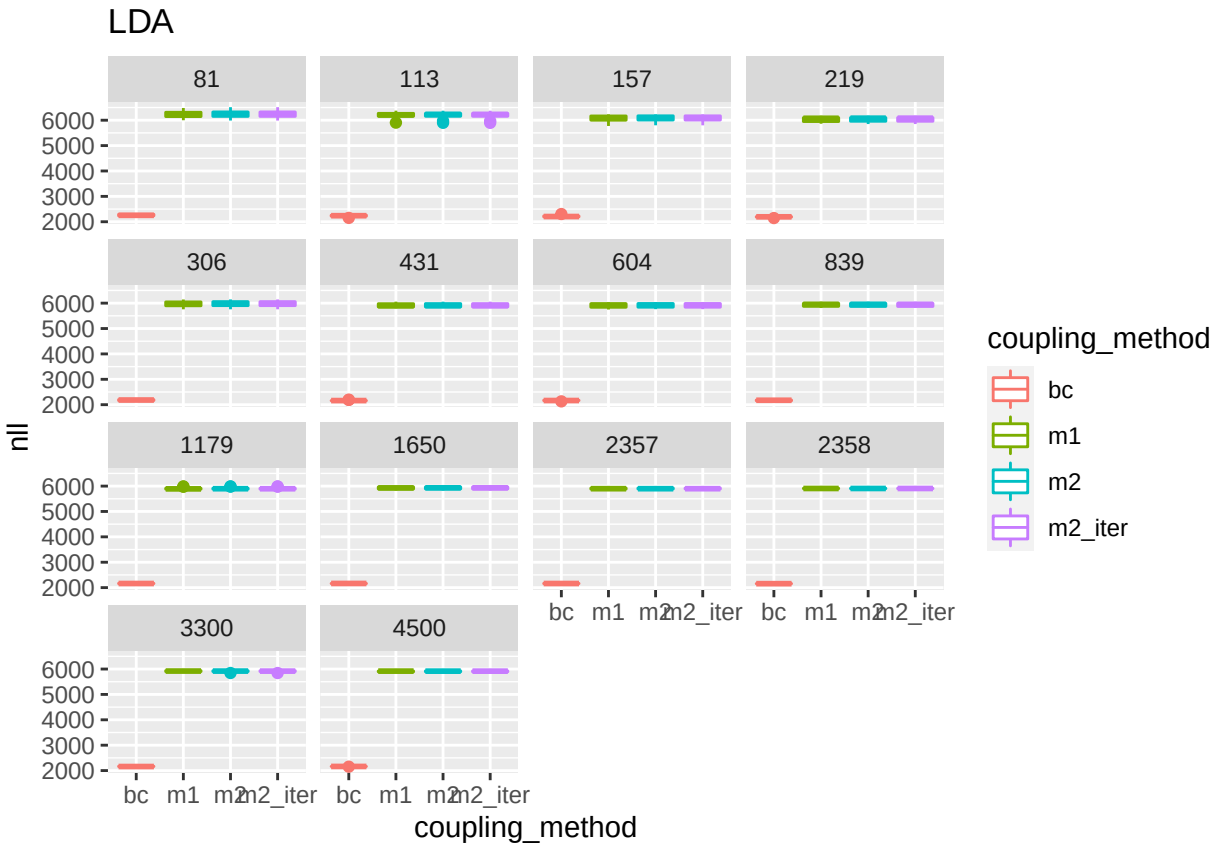
Here is visible some further increase, but it is not very stable and decreases again.

```
acc_ens_subsets %>% filter(1000 < train_size & train_size < 5000 & combining_method=="lda") %>% ggplot(
```



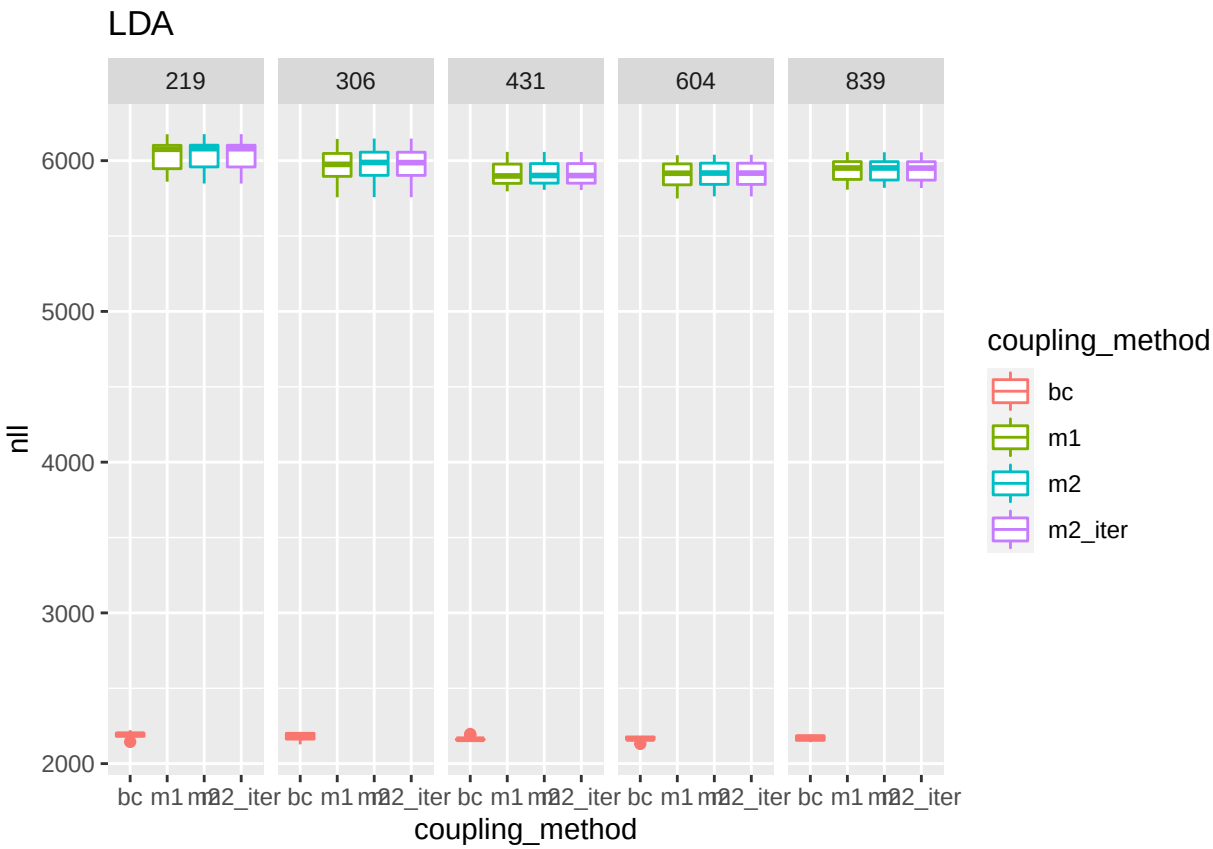
It seems, that from train set size around 1000 accuracy decreases again.

```
box_nll <- acc_ens_subsets %>% filter(combining_method=="lda") %>% ggplot() + geom_boxplot(mapping=aes(
box_nll
```



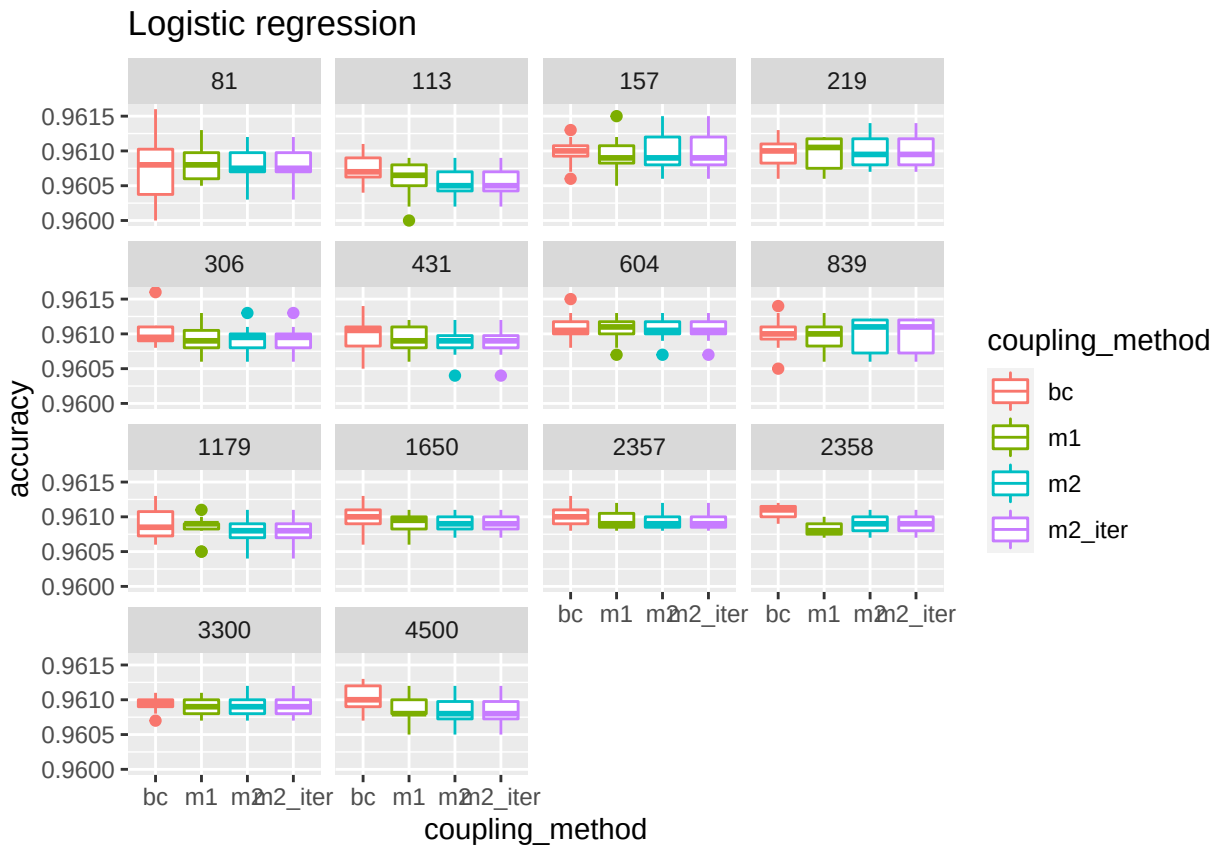
Nll doesn't seem to be doing much with changing LDA train size. Still, this metric needs further attention because of the zero probabilities produced by ensemble.

```
acc_ens_subsets %>% filter(200 < train_size & train_size < 1000 & combining_method=="lda") %>% ggplot()
```



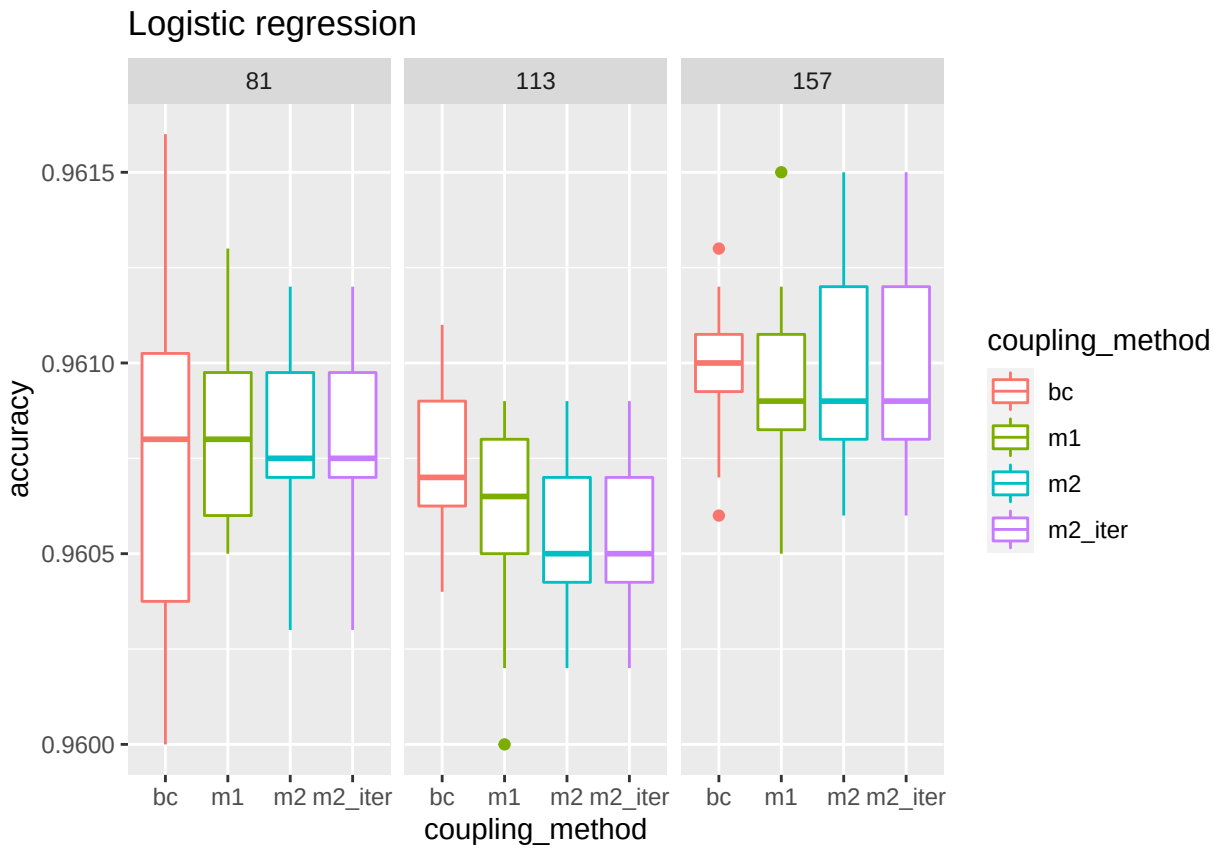
Logistic regression

```
box_acc <- acc_ens_subsets %>% filter(combining_method=="logreg") %>% ggplot() + geom_boxplot(mapping=ac
box_acc
```

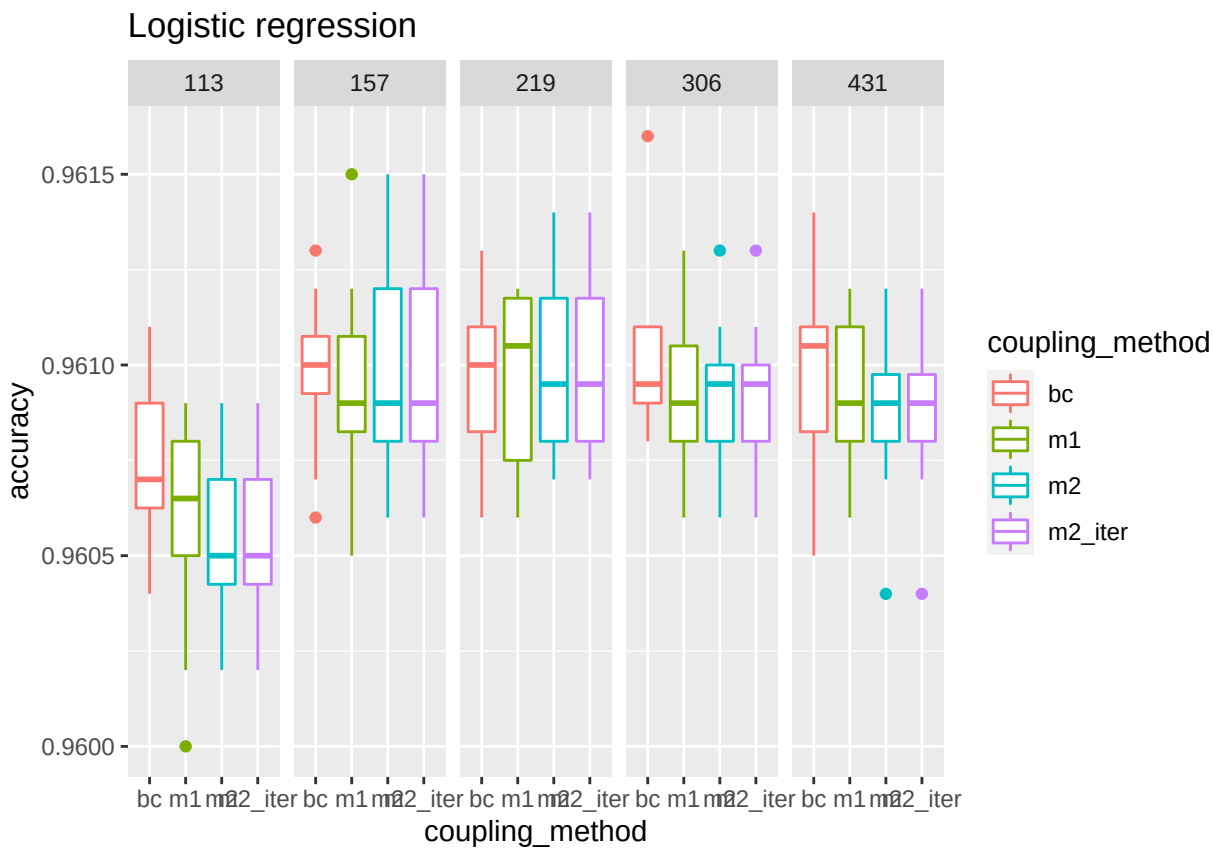


Accuracy seems to be slowly increasing with increasing training set size. We will inspect this closer by plotting for smaller subset sizes range.

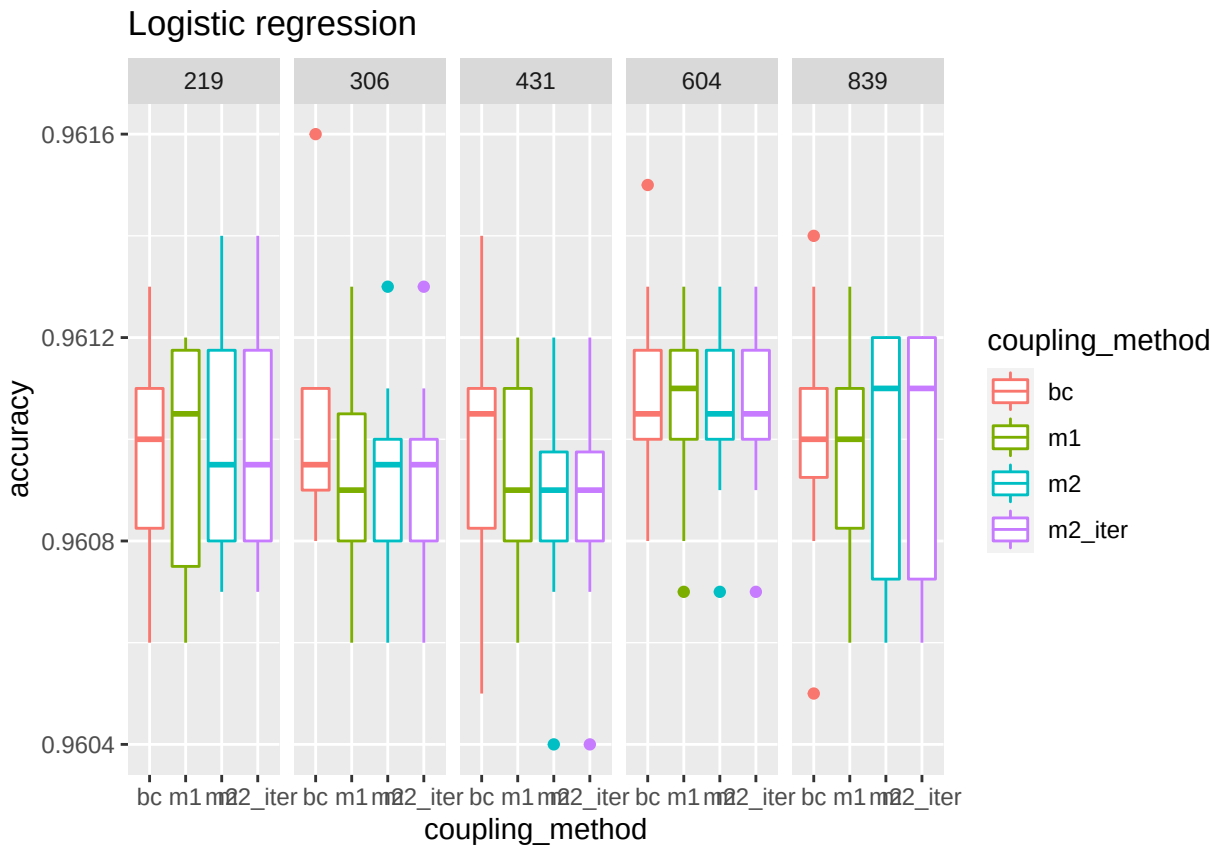
```
acc_ens_subsets %>% filter(0 < train_size & train_size < 200 & combining_method=="logreg") %>% ggplot()
```




```
acc_ens_subsets %>% filter(100 < train_size & train_size < 500 & combining_method=="logreg") %>% ggplot
```

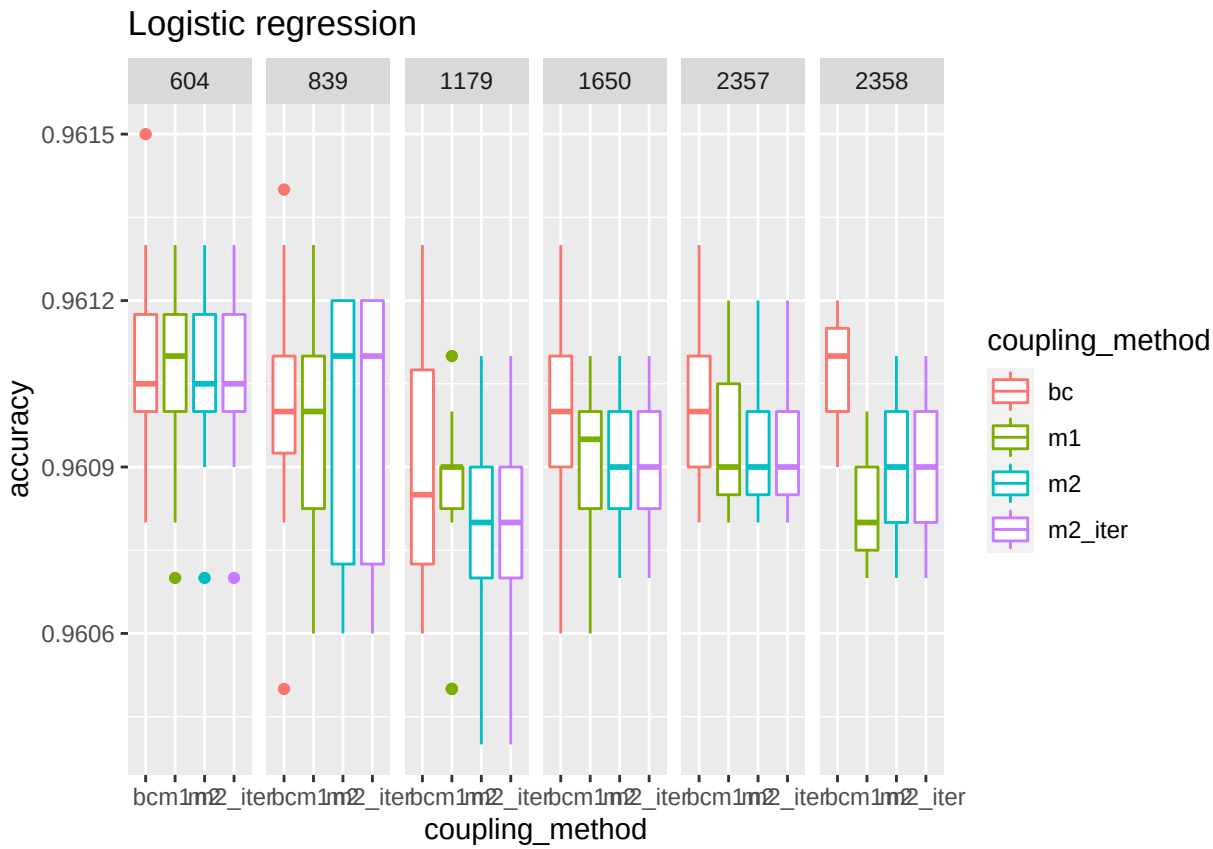



```
acc_ens_subsets %>% filter(200 < train_size & train_size < 1000 & combining_method=="logreg") %>% ggplo
```



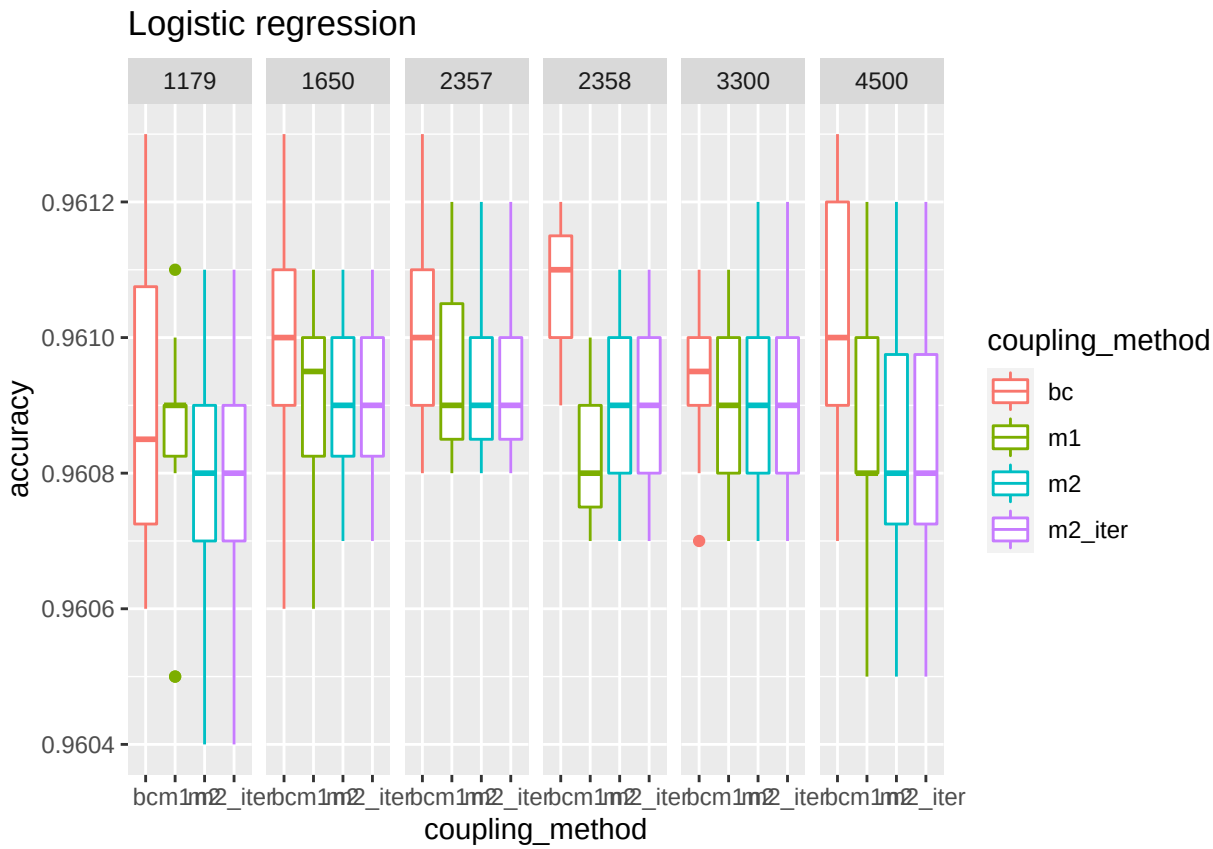
Here the accuracy increase seems to be stopping at train set size 604.

```
acc_ens_subsets %>% filter(500 < train_size & train_size < 3000 & combining_method=="logreg") %>% ggplo
```

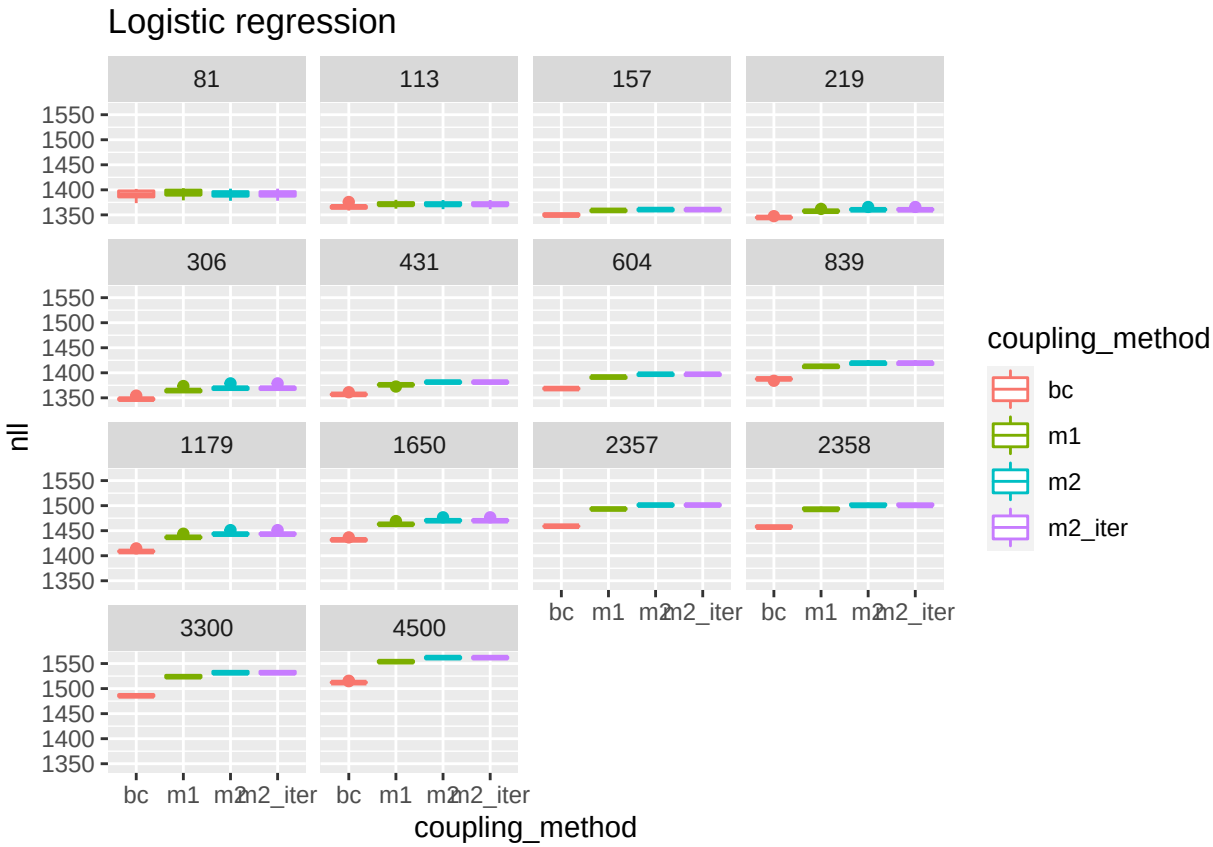


Here is visible some further increase, but it is not very stable and decreases again.

```
acc_ens_subsets %>% filter(1000 < train_size & train_size < 5000 & combining_method=="logreg") %>% ggpl
```

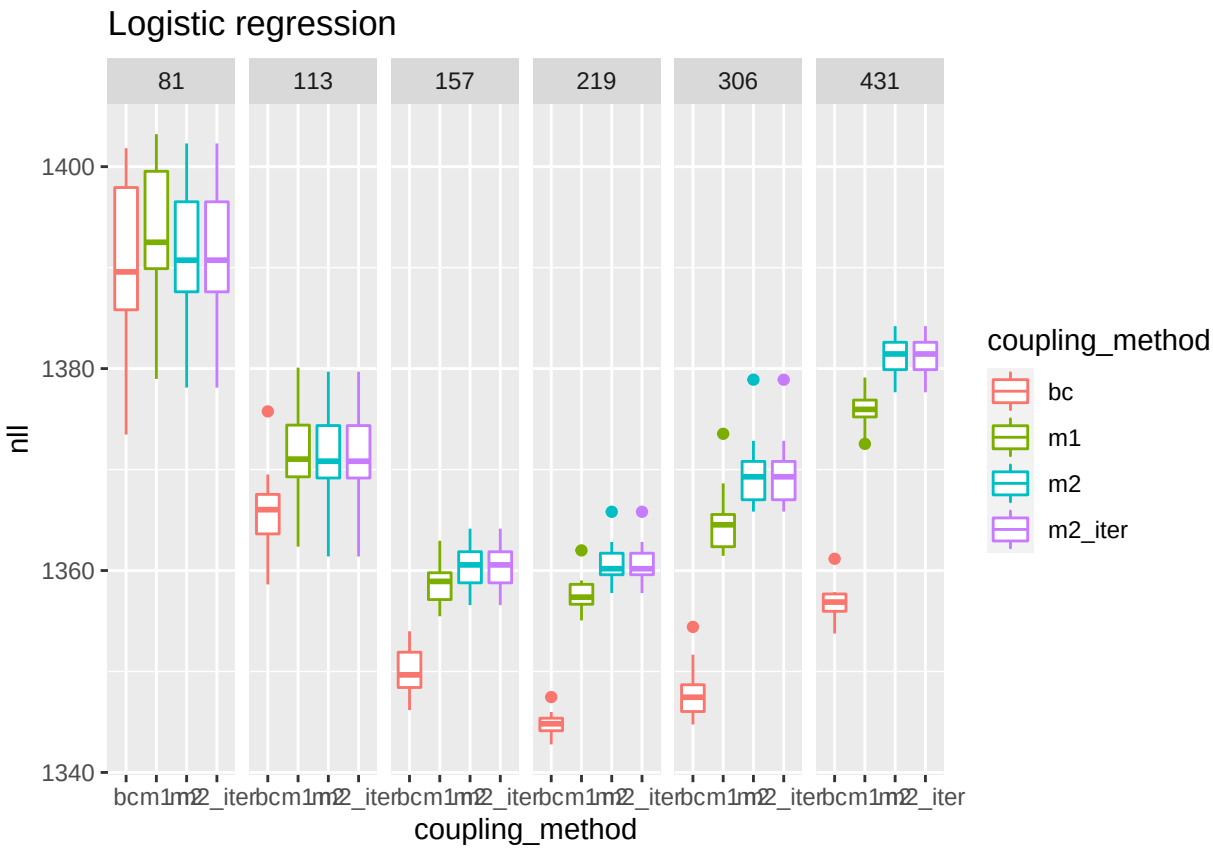



```
box_nll <- acc_ens_subsets %>% filter(combining_method=="logreg") %>% ggplot() + geom_boxplot(mapping=ac
box_nll
```



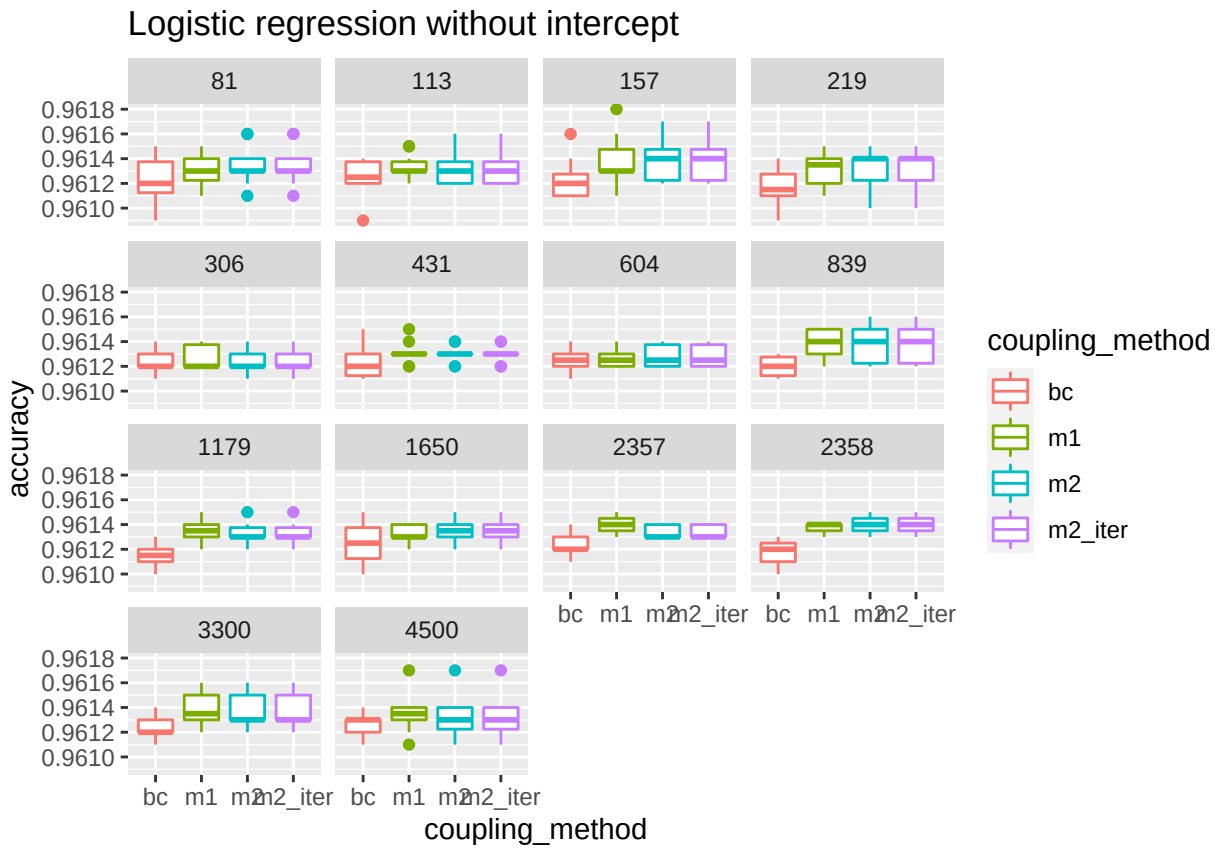
Nll is decreasing with increasing train set size for up to size around 200, then it starts increasing.

```
acc_ens_subsets %>% filter(0 < train_size & train_size < 500 & combining_method=="logreg") %>% ggplot()
```



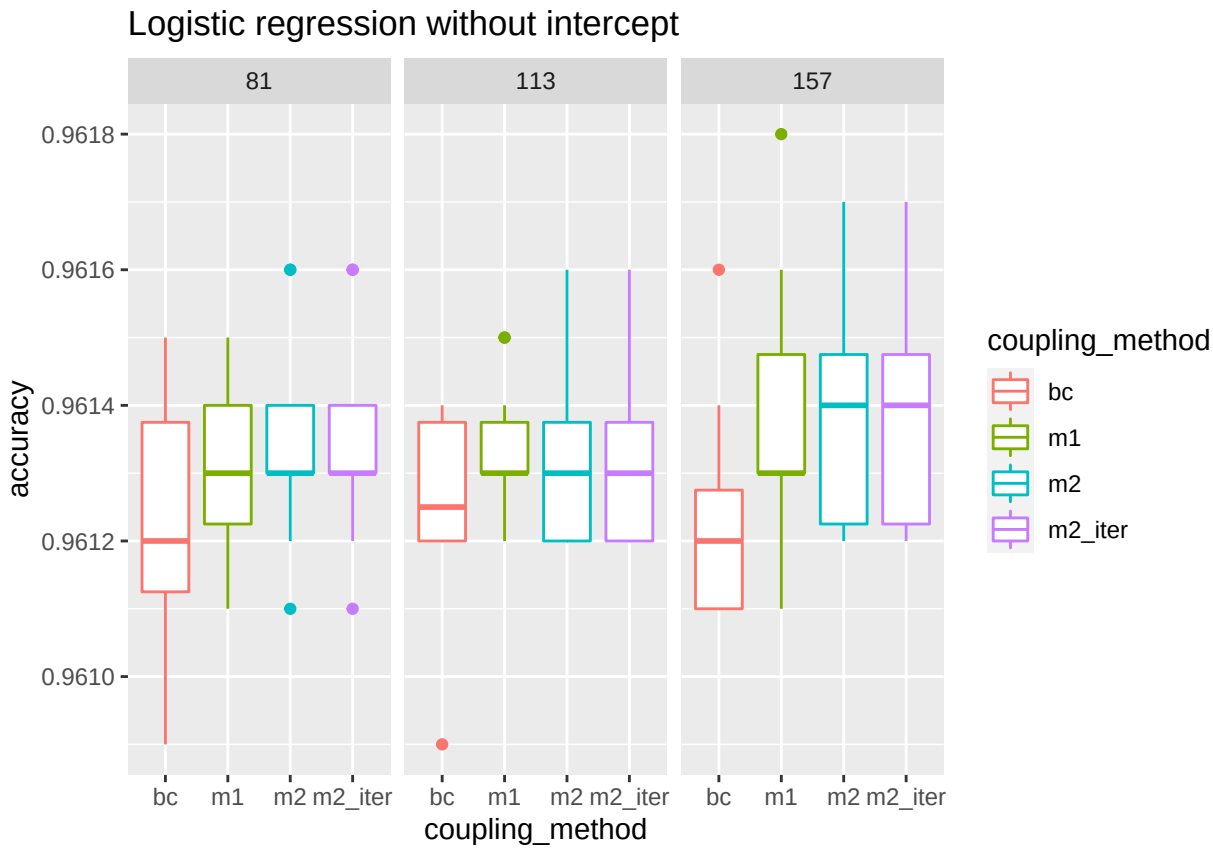
Logistic regression without intercept

```
box_acc <- acc_ens_subsets %>% filter(combining_method=="logreg_no_interc") %>% ggplot() + geom_boxplot
box_acc
```

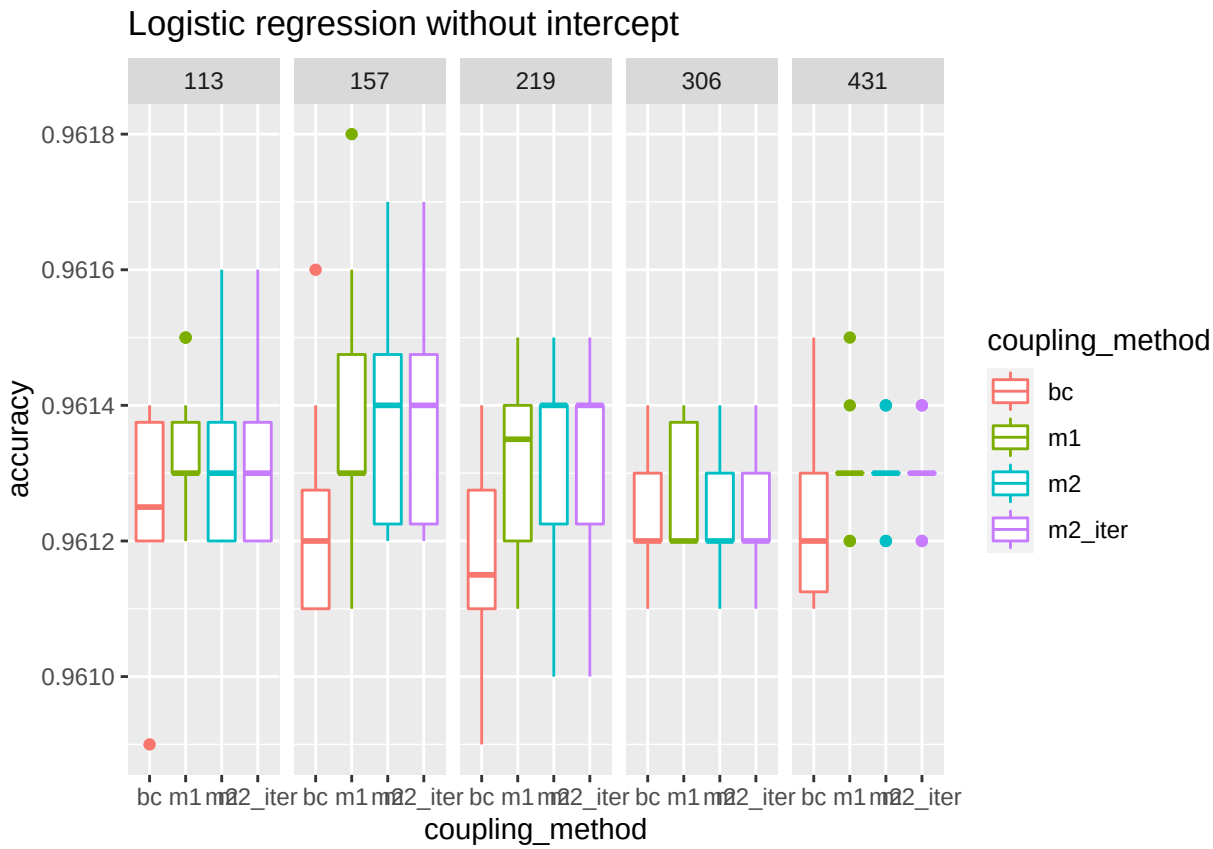


Accuracy seems to be slowly increasing with increasing training set size. Coupling method bc achieves inferior results for some train sizes. We will inspect this closer by plotting for smaller subset sizes range.

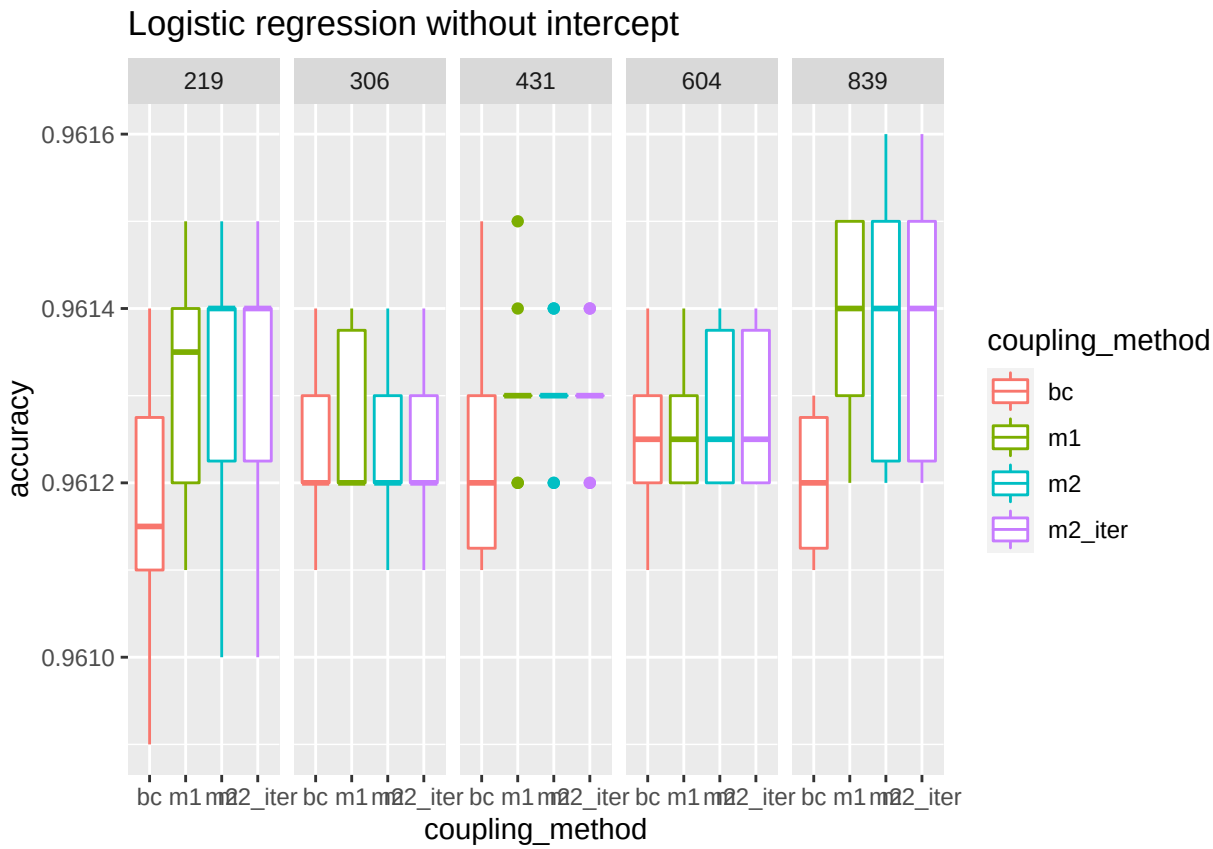
```
acc_ens_subsets %>% filter(0 < train_size & train_size < 200 & combining_method=="logreg_no_interc") %>%
```



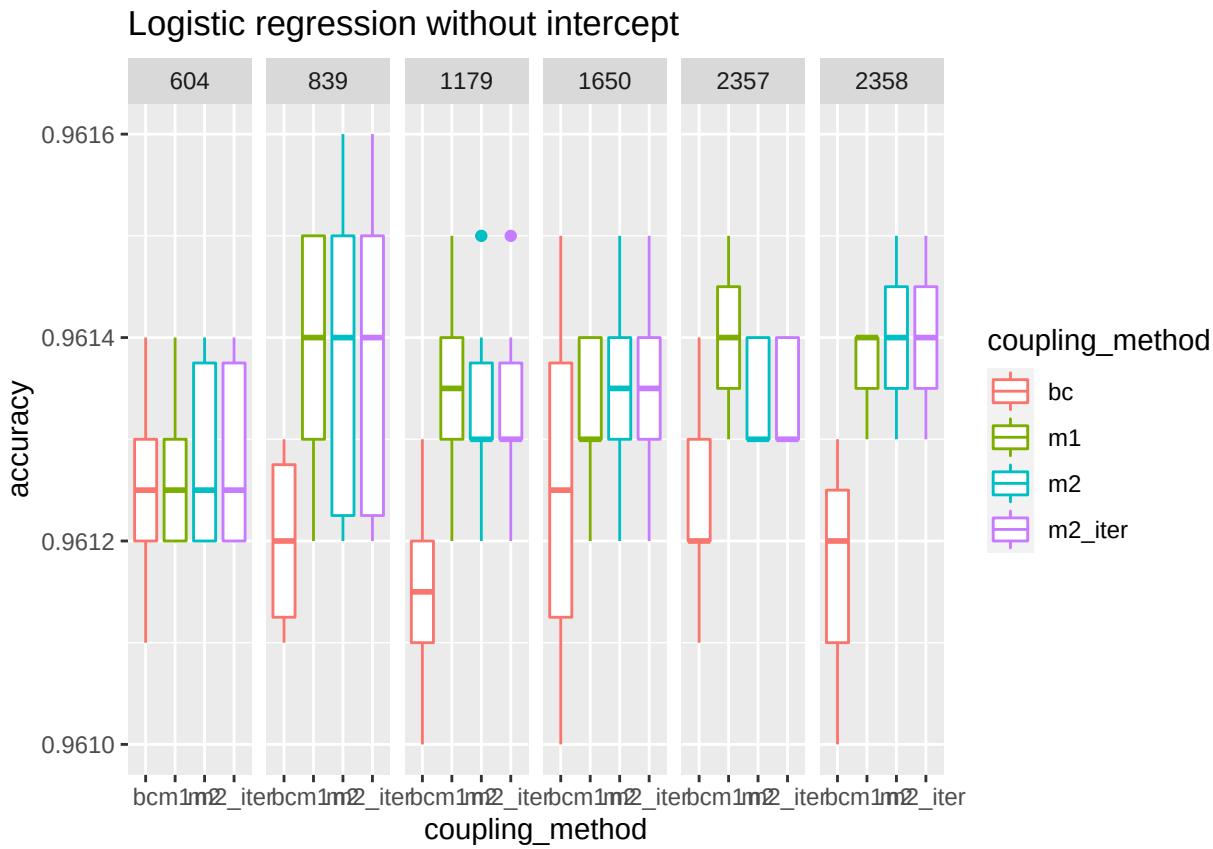

```
acc_ens_subsets %>% filter(100 < train_size & train_size < 500 & combining_method=="logreg_no_interc") %
```




```
acc_ens_subsets %>% filter(200 < train_size & train_size < 1000 & combining_method=="logreg_no_interc")
```

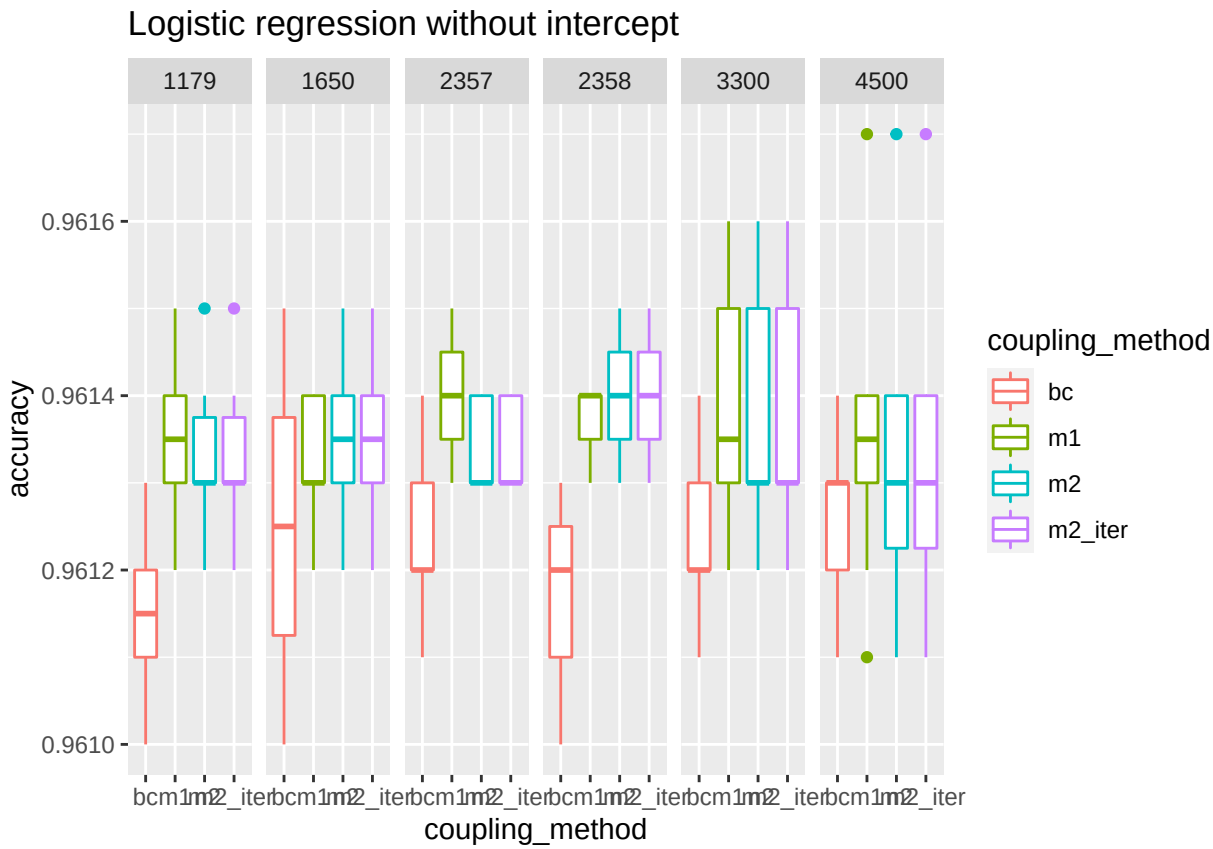



```
acc_ens_subsets %>% filter(500 < train_size & train_size < 3000 & combining_method=="logreg_no_interc")
```



Here the accuracy increase seems to be stopping at train set size 839. However, accuracy of method bc is already decreasing.

```
acc_ens_subsets %>% filter(1000 < train_size & train_size < 5000 & combining_method=="logreg_no_interc"
```



It seems, that from train set size around 1000 accuracy is stagnating.

```
box_nll <- acc_ens_subsets %>% filter(combining_method=="logreg_no_interc") %>% ggplot() + geom_boxplot
box_nll
```

Nll is decreasing with increasing train set size for up to size around 200, then it starts increasing.

```
acc_ens_subsets %>% filter(0 < train_size & train_size < 500 & combining_method=="logreg_no_interc") %>%
```

In LDA and logistic regression with intercept, good tradeoff between train set size and good accuracy seems to be around 500. For logistic regression without intercept around 800. Negative log likelihood for lda seems mostly unaffected by changing train set size, for logistic regression it is minimal at around train set size 200.

Šuch, Ondrej, and Santiago Barreda. 2016. "Bayes Covariant Multi-Class Classification." *Pattern Recognition Letters* 84: 99–106.

Wu, Ting-Fan, Chih-Jen Lin, and Ruby C Weng. 2004. "Probability Estimates for Multi-Class Classification by Pairwise Coupling." *Journal of Machine Learning Research* 5 (Aug): 975–1005.