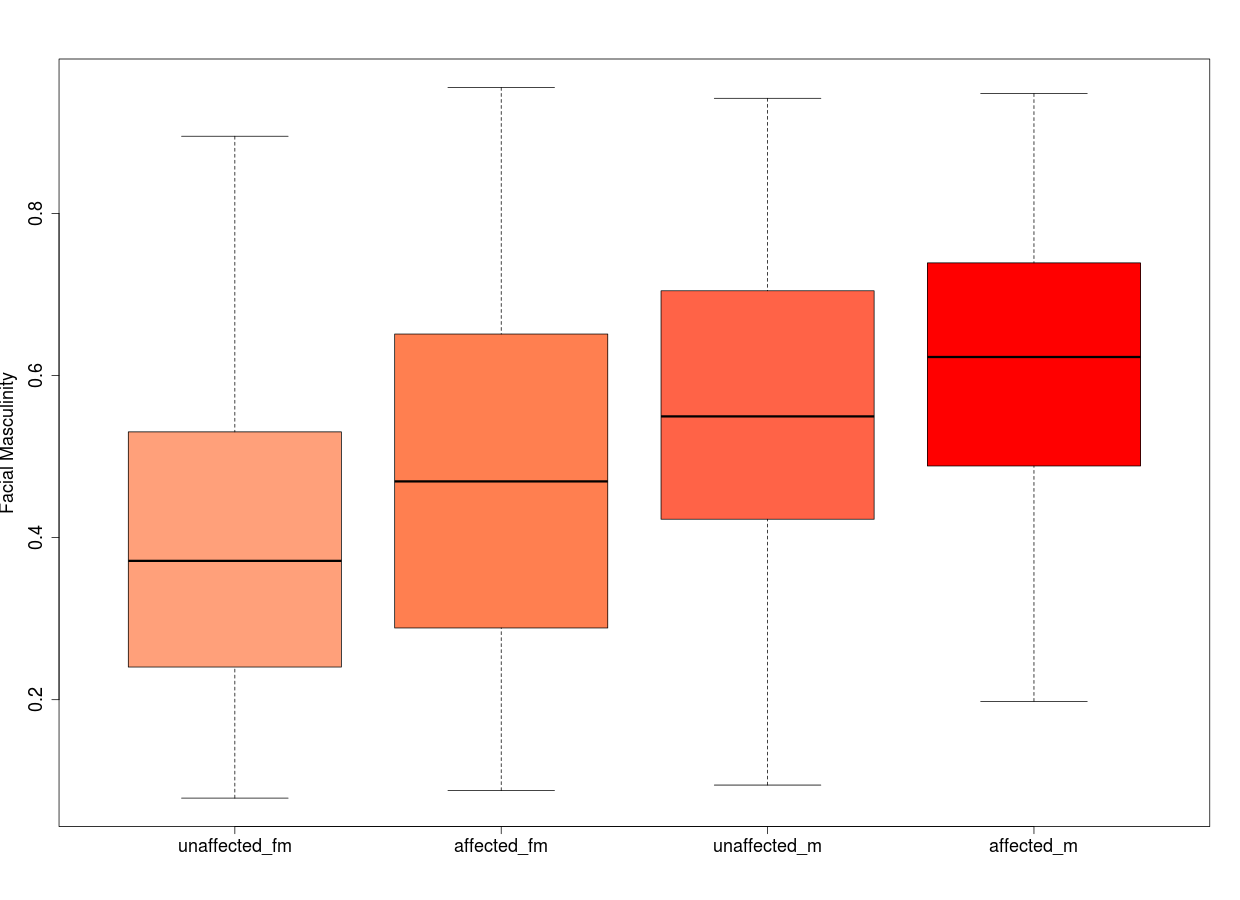
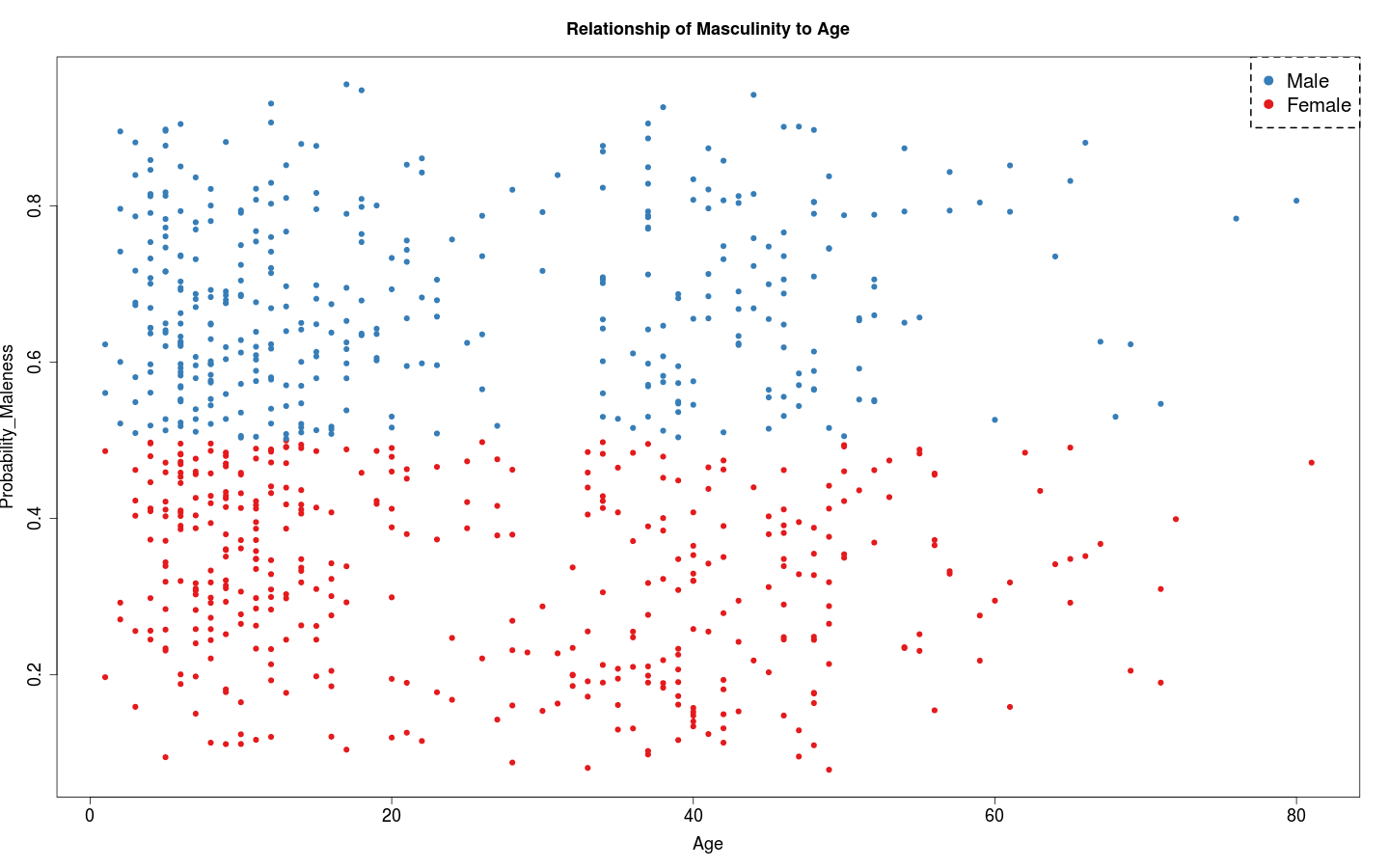
**Machine learning approach to predict “maleness”(svm)**

(/wdata/rotating\_students/yonghuang/data/svm\_test.R)

*Figure 1.1. This figure shows the relationship of facial masculinity to participant's sex and affected status. X axis is labeled as different sex and affected status. Y axis is labeled as the probability of maleness.*

After 10-fold cross-validation training on **svm model**, the result shows that both affected male and female tend to have higher masculinity than unaffected male and female. Also, affected male and female have higher masculinity than unaffected male and female.

Besides, the accuracy of this model is between 65% to 71% under 97.5% confidence interval.



*Figure 1.2. The figure shows the probability of maleness from 772 participants based on their age.*

*Red dot indicates the female and Blue dot indicates male.*

Based on 10-fold cross validation training on svm model, the prediction results from training interprets distinct difference of masculinity between male and female in all age groups. However, the data is evenly distributed among all age groups. It does **not show a linear decline or increase** of masculinity distribution in different age groups.

lm(formula = Probability\_maleness ~ Sex + Age, data = sub\_features)

Residuals:

Min 1Q Median 3Q Max

-0.243123 -0.094832 -0.002657 0.097327 0.275485

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.0232930 0.0146336 -1.592 0.112

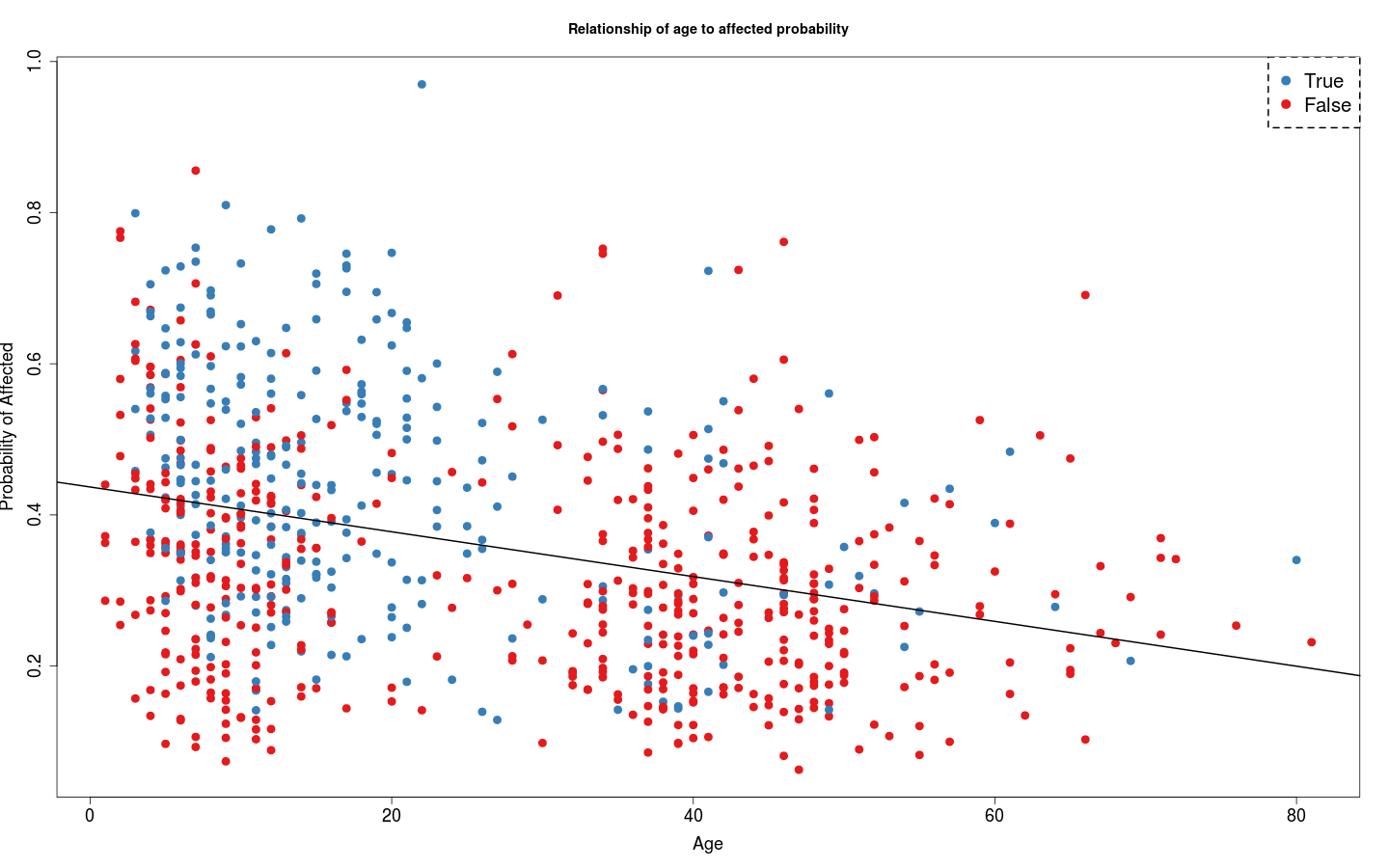
Sex 0.3530983 0.0083331 42.373 <2e-16 \*\*\*

Age -0.0001726 0.0002291 -0.753 0.452

We test the significant of probability of maleness to Sex and Age from a linear model. The statistic result above shows that **Sex has a pValue smaller than 0.05**, which indicates that Sex is significantly related to the probability of maleness. Besides, the **Age has a pValue greater** than 0.05, which indicates that **age is not a confounding factor** to associate with probability of maleness.

**Machine learning approach to test “Affected Status” (svm)**

(/wdata/rotating\_students/yonghuang/data/Affected\_status(svm).R)



*Figure 1.2. The figure shows the probability of autism from 772 participants based on their age.*

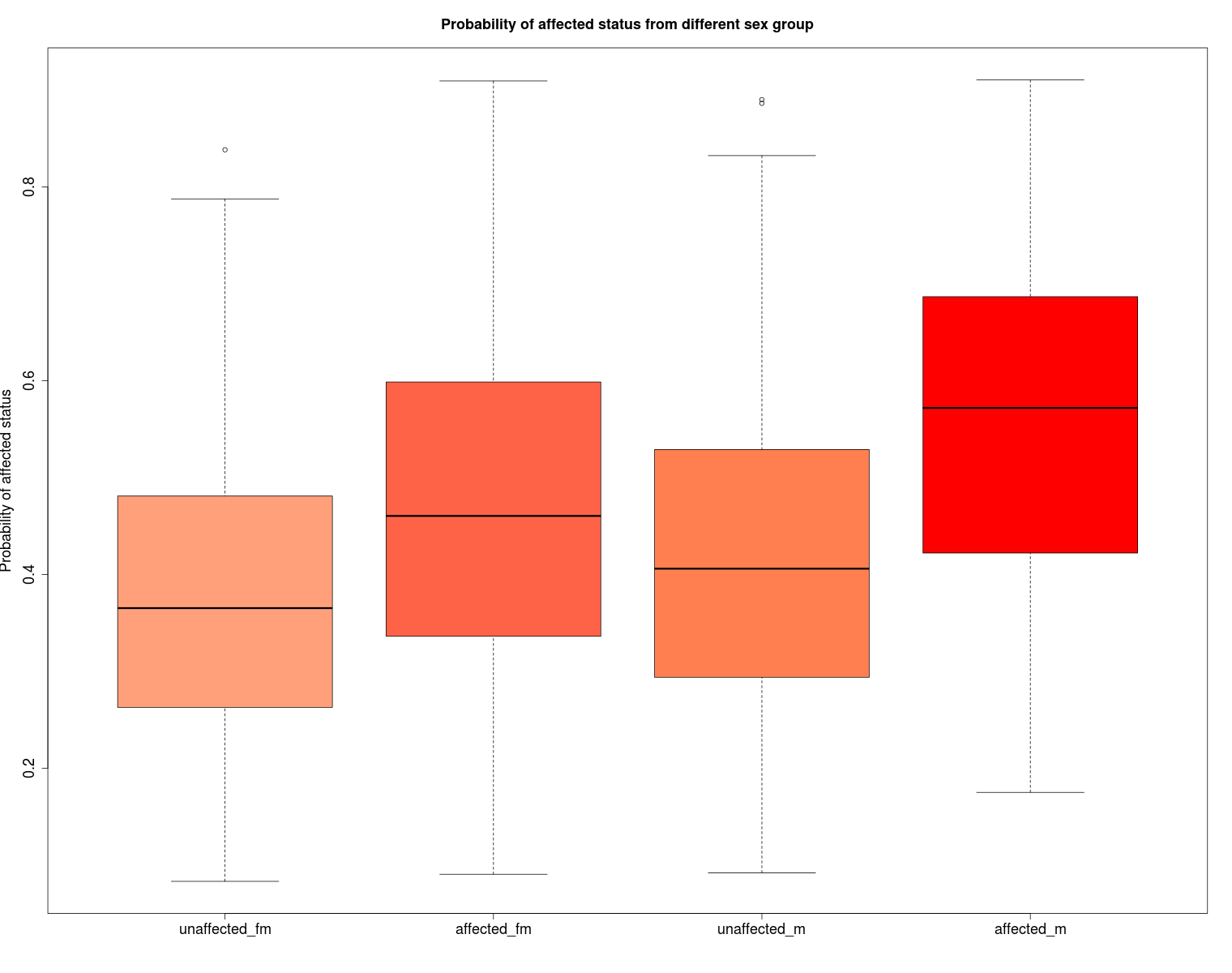
*Red dot indicates the false and Blue dot indicates true.*

The model is similar to test for maleness. The figure above represents **a decreasing** probability of autism through **increasing** age in overall distribution. Also, the **blue dots,** which represent the true prediction label for autism, are centralized around **0 to 30 years old** group. The **red dots** are centralized around **0 to 20** and **30 to 50** age groups.

Since the we are unable to **stratify** the svm model, the data are unbalanced between true and false labels. The overall result for predicting the autism contains bias. Also, the during the cross-validation, the svm model **would flip the label class**, so it is hard to format the overall data for prediction. (the above graph is adjusted and does not have this problem now)

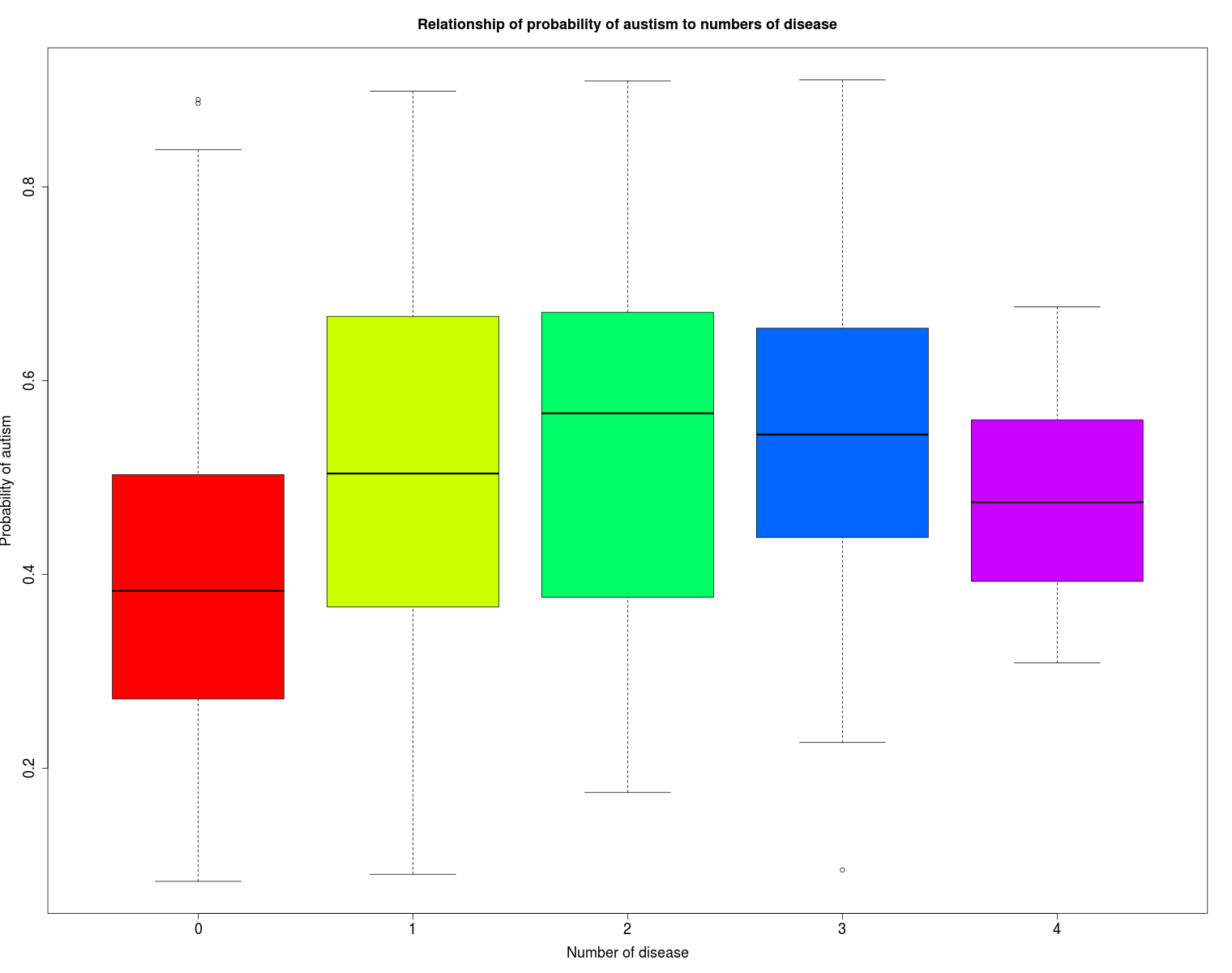
**Machine learning approach to test “Affected Status” (Random Forest)**

(/wdata/rotating\_students/yonghuang/data/RF\_Affected\_status.R)

 *Figure 2.1 This figure indicates the relationship of Sex to affected probability of autism.*

The average affected probability of autism is higher for both female and male. In the contrary, the female and male who are not affected tend to have lower affected probability of autism compared to affected female and male. In the other hand, in unaffected group, average affected probability of male is higher than female. Also, in affected group, average affected probability of male is higher than female.

Overall, the figure indicates that **affected people** would have **higher probability** than normal people. Also, **male** tends to have **higher affected probability** of autism than **female.**



*Figure 2.2 The figure represents the association of number of disease to the probability of autism.*

The red box is the group of people (489) who did not have any disease and **has lowest probability.** The yellow box is the group of people(128) who has only 1 disease. The green box is the group of people(91) who has 2 diseases and **has highest probability**. The blue box is the group of people(57) who has 3 diseases. The purple box is the group of people(7) who has 4 diseases.

The figure indicates that people have **no disease** related to autism would have **lowest probability**. Besides, increasing number of disease implies increase probability of autism from 0 to 2. Then, the probability decrease slightly from 3 to 4.

Here is the linear model result from associating the affected probability to 4 different disease.

lm(formula = rf$votes[, "TRUE"] ~ ID + Autism.reported. + Language.impair + Epilepsy.status, data = subfeatures)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.403282 0.007182 56.152 < 2e-16 \*\*\*

IDTRUE 0.002083 0.019012 0.110 0.91279

Autism.reported.TRUE 0.115063 0.015726 7.317 6.41e-13 \*\*\*

Language.impairTRUE 0.058329 0.017668 3.301 0.00101 \*\*

Epilepsy.statusTRUE 0.023505 0.038847 0.605 0.54532

The **pValue** of reported autism cases and language impairment show **great significant** to prediction of autism. The **pValue** of epilepsy and intellectual disorder show **no significance** to prediction of autism.

**The linear model result on different associations of diseases.**

**A> lm(formula = Epilepsy.status ~ rf$votes[, "TRUE"] + Autism.reported. + Language.impair + ID, data = subfeatures)**

Residuals:

Min 1Q Median 3Q Max

-0.13554 -0.02205 -0.00673 -0.00312 0.99226

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.002759 0.015088 -0.183 0.855

rf$votes[, "TRUE"] 0.020297 0.033546 0.605 0.545

Autism.reported.TRUE 0.024581 0.015089 1.629 0.104

Language.impairTRUE -0.017083 0.016523 -1.034 0.302

IDTRUE 0.096608 0.017319 5.578 3.37e-08 \*\*\*

We associate the Epilepsy with affected probability, reported autism status, language impairment and intellectual disorder in linear model. The statistic results shows that **only intellectual disorder** has **pValue smaller** than 0.05, which indicates that epilepsy is not significantly associate with affected probability of autism and language impairment.

----------------------------------------------------------------------------------------------------------------------------

**B> lm(formula = Autism.reported. ~ rf$votes[, "TRUE"] + Epilepsy.status + Language.impair + ID, data = subfeatures)**

Residuals:

Min 1Q Median 3Q Max

-0.92213 -0.19498 -0.08553 0.03183 1.05924

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.11057 0.03582 -3.087 0.0021 \*\*

rf$votes[, "TRUE"] 0.56704 0.07750 7.317 6.41e-13 \*\*\*

Epilepsy.statusTRUE 0.14028 0.08611 1.629 0.1037

Language.impairTRUE 0.23240 0.03860 6.021 2.68e-09 \*\*\*

IDTRUE 0.20734 0.04154 4.992 7.40e-07 \*\*\*

We associate the reported autism status with affected probability, epilepsy, language impairment and intellectual disorder. The statistic results show that **only epilepsy** has **pValue greater** than 0.05, which indicates that epilepsy is not significantly associate with reported autisim status.

----------------------------------------------------------------------------------------------------------------------------

**C> lm(formula = Language.impair ~ rf$votes[, "TRUE"] + Autism.reported. + Epilepsy.status + ID, data = subfeatures)**

Residuals:

Min 1Q Median 3Q Max

-0.85441 -0.11359 -0.06591 -0.02235 0.98611

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.01927 0.03294 -0.585 0.55866

rf$votes[, "TRUE"] 0.24021 0.07276 3.301 0.00101 \*\*

Autism.reported.TRUE 0.19421 0.03225 6.021 2.68e-09 \*\*\*

Epilepsy.statusTRUE -0.08147 0.07880 -1.034 0.30150

IDTRUE 0.47698 0.03452 13.816 < 2e-16 \*\*\*

We associate the language impairment with affected probability, reported autism status, intellectual disorder. The statistic results indicate that **epilepsy** has **pValue greater** than 0.05, which means that epilepsy is not significant factor associated with language impairment.

----------------------------------------------------------------------------------------------------------------------------

**D> lm(formula = ID ~ rf$votes[, "TRUE"] + Autism.reported. + Language.impair + Epilepsy.status, data = subfeatures)**

Residuals:

Min 1Q Median 3Q Max

-0.61893 -0.04703 -0.04543 -0.04413 0.95640

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.042551 0.030798 1.382 0.167

rf$votes[, "TRUE"] 0.007513 0.068576 0.110 0.913

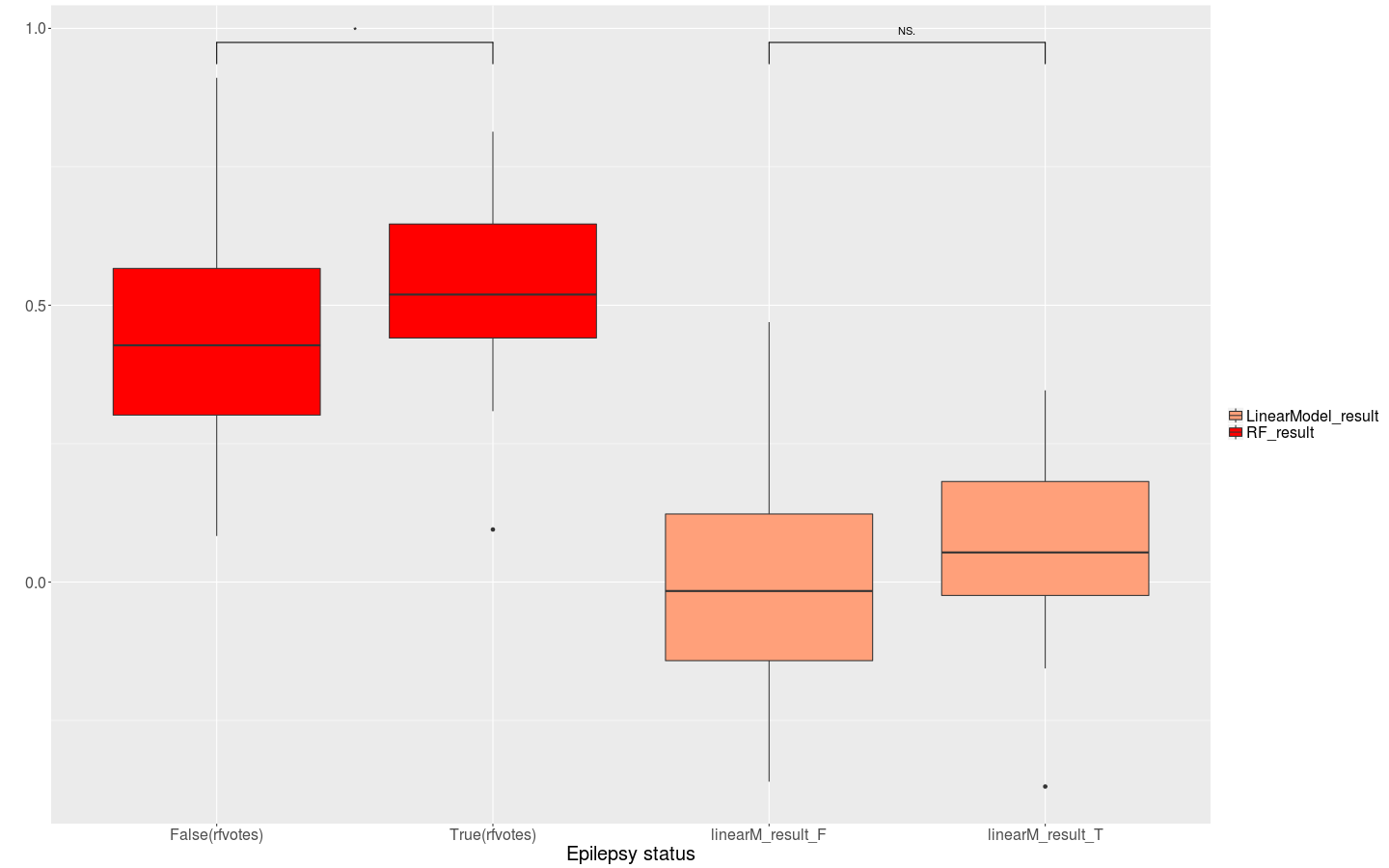
Autism.reported.TRUE 0.151761 0.030402 4.992 7.40e-07 \*\*\*

Language.impairTRUE 0.417788 0.030239 13.816 < 2e-16 \*\*\*

Epilepsy.statusTRUE 0.403542 0.072344 5.578 3.37e-08 \*\*\*

We associate the intellectual disorder with reported autism status, intellectual disorder, language impairment and epilepsy. The statistic results show that **pValue** of **reported autism status**, l**anguage impairment and epilepsy** is **smaller** than 0.05, which indicates that **intellectual disorder** is a **significant factor** associated with reported autism status, language impairment and epilepsy.

**Epilepsy specific test result**

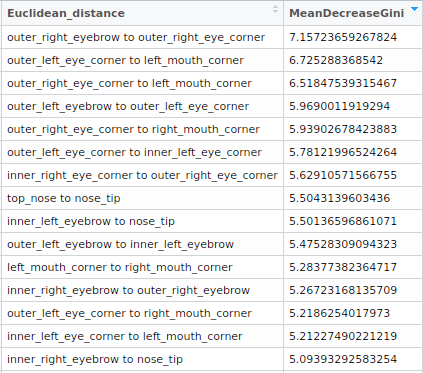
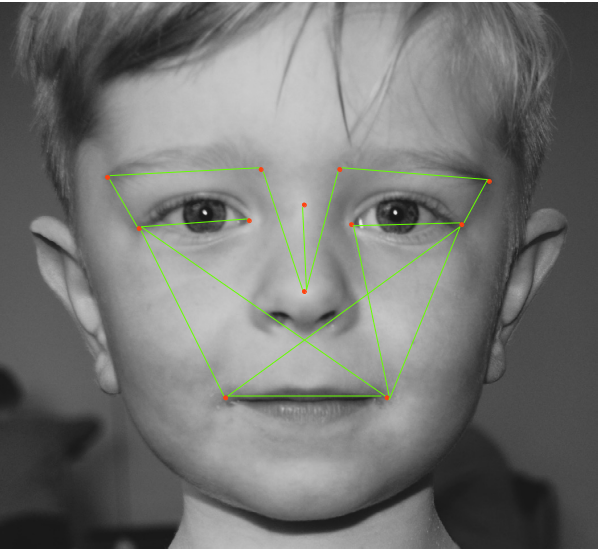
*Figure 2.3 Comparison of epilepsy status result between affected probability from random forest training(Red boxes) and the residues from linear model that eliminate the effect of confounding factors(Light Salmon).*

The two red plots on the left ,**including all the confounding factors**, have a start sign above that indicates greater and significant difference between true and false value, which illustrate that participants who have **epilepsy have higher probability of autism**.

However, the two salmon color plots on the right is constructed after **eliminating the effect of reported cases of autism, language disorder and intellectual disorder**. Then, the NS sign above indicates the true and false distribution of **epilepsy** is similar and **not significant** to predict the probability of autism. Furthermore, the average residue values are approximate 0, which further illustrates that **epilepsy does not have influence** on predicting the autism.

Overall, from the linear model results and box plot above, the reported cases of autism and language impairment shows significant to the affected probability of autism. In the other hand, after eliminating the confounding factors in the linear model test, epilepsy is not significant to the prediction of autism.

**Facial feature detection result**

 *Table 1.1 Euclidean distance importance*  *Figure 2.4 Facial feature representation based on table 1.1*

This table represents top 15 important features out of 66 related to the affected probability of autism after training the dataset from random forest approach. In specifically, the “outer\_right\_eyebrow to outer\_right\_eye\_corner” has importance of 7.157, which is the most important feature in the dataset.

Based on the table, the figure in the right shows where those features located in actual face. Besides, the combination of green lines indicates that those features is almost symmetric.

**The confusion matrix from svm and random forest model in affected status analysis**

Random forest

prediction\_label

FALSE TRUE

FALSE 341 148

TRUE 111 172

SVM

prediction\_label

FALSE TRUE

FALSE 416 73

TRUE 164 119

From the two matrices above, the random forest performs better at predicting the true positive label for autism than svm. However, neither random forest and svm perform good at predicting the true positive result.

Deep learning

prediction\_label

False True

False 382 107

True 130 153

threshold specificity sensitivity (CNN)

0.3267695 76.8916155 57.5971731

threshold specificity sensitivity (RF)

0.5422379 82.2085890 49.4699647

DeLong's test for two ROC curves

data: rocobj1 and rocobj2

D = 0.053763, df = 1541.8, p-value = 0.9571

alternative hypothesis: true difference in AUC is not equal to 0

sample estimates:

AUC of roc1 AUC of roc2

71.02943 70.88166