Exploratory Analysis of

US Lung Disease in Chronic Airway Infections

* Summary

We want to know: is there any relation between explanatory variables and response variable. In this case, we aim to detect the relation between expiratory volume among the patient who has lung disease and other possible variables. And if we could fit a model for the data, how it looks like. First, we need to figure out which variables will be included in the model by using a stepAIC function in R. And then test the efficiency of explanatory variables. Also, we need to test whether the target model meets the requirement of the model hypothesis. Second, try to include some interaction terms or quadratic forms. Then, see model comparison results from ANOVA function and diagnostic plots in R, to determine which model we will remain. So, based on the above methods, the final model will use multiple linear regression model that will be(s are coefficients):

* Introduction

With rest, supportive care and occasional antibiotics, most lung infections should improve in a few weeks. But if your symptoms persist beyond this time, you may have a chronic pulmonary infection.1

|  |  |  |  |
| --- | --- | --- | --- |
| The symptoms of chronic respiratory infections include2: | | | |
| Shortness of breath | Fatigue | Mucus production | Fever |
| Sore throat | Postnasal drip or nasal discharge | Bad breath | Cough |

|  |  |  |
| --- | --- | --- |
| Treatment include2: | | |
| Techniques to clear mucus | Antibiotics | Inhaled medicine |
| Smoke avoidance | Antiviral medication |  |

Due to the above symptoms, the data is designed to measure the expiratory volume as the levels of chronic respiratory infections effect.

Because linear regression model uses ordinary least squares method to estimate the coefficient of variables. So if we want to use the linear regression model, we have to meet the requirement of hypothesis of ordinary least squares method. They are: errors are independent, have equal variance and follow normal distribution. Also, we need to detect multicollinearity among explanatory variables. If multicollinearity exist, the [coefficient estimates](https://en.wikipedia.org/wiki/Regression_coefficient) of the multiple regression may change erratically in response to small changes3. So it will hurt the accuracy of coefficients.

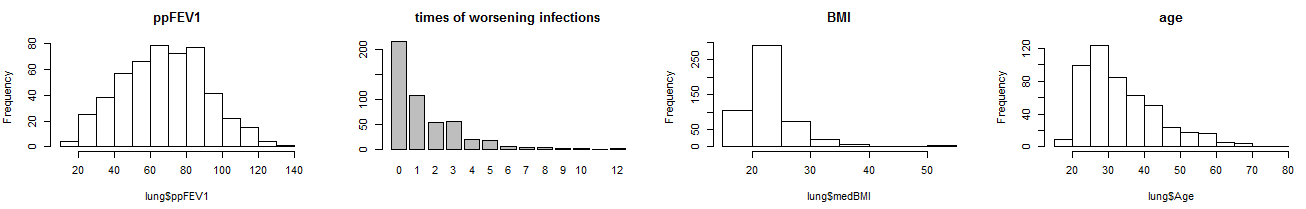
Here, I use both ANOVA and AIC function in R to choose a proper model. If the p-value between two models in the ANOVA result is smaller than 0.05, I will keep the full model. It means two models are significantly different. Meanwhile, Comparing two AIC values, pick up the model which has a smaller AIC value. Because of the smaller the AIC, the better the fit. But sometimes the two models have kind of similar AIC value, then I will see the ANOVA result, if p-value is greater than 0.05, I will keep the reduced model, just want to have a simple model with more efficiency. Also, I will see diagnostic plot of the fit, to check whether the errors meet the requirement of hypothesis. Other details of model selection processes state in the following chapters.

* Data description

The raw data is a sample from patients diagnosed with lung disease associated with chronic airway infections in the U.S. during 2014. There are 500 observations on 23 variables. The variable values from one observation represent one patient's summary information for the year. I choose ppFEV1 as response variable, the meaning of it is: how much expiratory volume the patients with lung disease achieved. The explanatory variables are: times of worsening of airway infections, height, weight, BMI, age, gender, 4 types of races, enzymes, 5 types of bacteria, 3 types of therapies, exercise, depression, and diabetes. For complete explanation of each variable, see Appendix part2.

Now, we need variable selection. I use stepwise AIC model selection method. First, fit the data from linear regression model. Then, use "step" function in R get the result. See output in Appendix output1. Finally, the selected variables are: PulmonaryExacerbations, medBMI, Age, Gender, PA, MRSA, TOBI, DailyAirway, RegExercise, Depression, and Diabetes.

The figure1 are histograms of response variable and selected variables. The range of numeric variables are: PulmonaryExacerbations [0,12], medBMI [15.29401,52.58919] and Age[19,79].



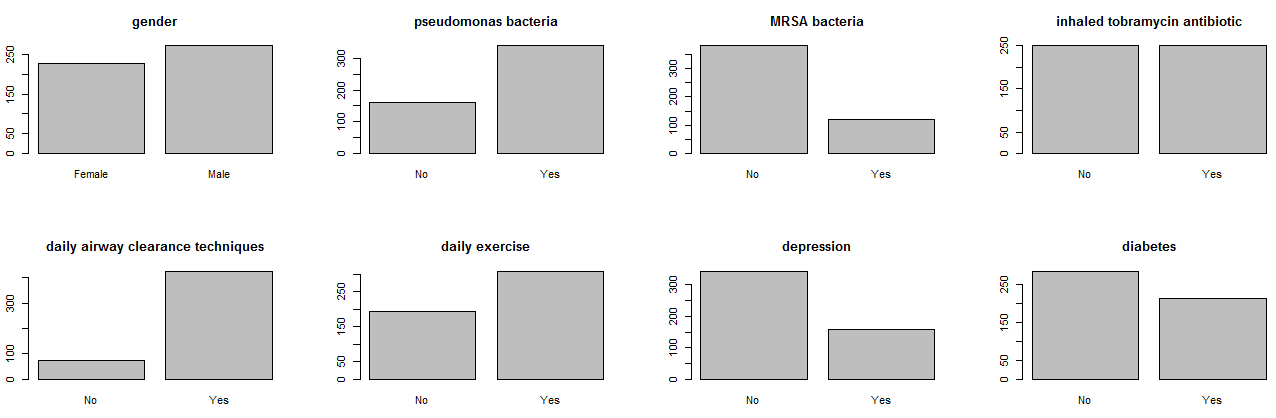


figure1

* Statistical analysis

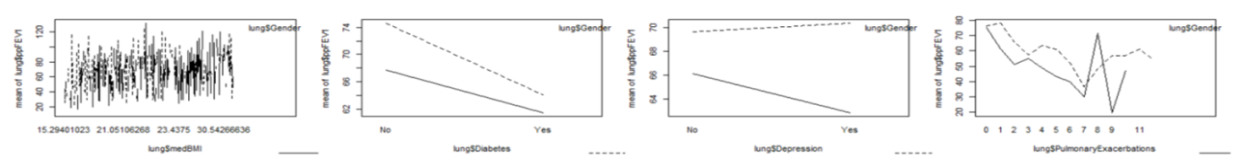
All the following fit number and model details are in Appendix part3.

I try to use log transformation on ppFEV1, to see any improvement. the AIC value of the model after log transformation is way better than without log transformation. But when I check the diagnostic plot of log transformation model, it shows obviously some violence of normality of errors in the two tails. Also, the histogram of log transformation model shows the ppFEV1 are clearly skewed. So, I keep fit3, ignore the log transformation form due to this reason. See Appendix output2.

Even the stepwise AIC give the variable selection result, I also want to double check by using ANOVA method. See Appendix output3. Two coefficient p-values are greater than 0.05. Depression's p-value is 0.126060 and Diabetes's p-value is 0.118295. Comparing two models with ANOVA, p-value is 0.09193. ANOVA function uses likelihood ratio test to detect the difference between two models.

## Here, you have two choices: keep simple model due to ANOVA or keep full model due to AIC. AIC means Akaike's An Information Criterion. StepAIC means exact AIC. They have different algorithm. Comparing two models I just use AIC. I will keep these two terms now since I have some other model selection tests that are related to them.

I want to test any interaction terms exists. First: interactions among gender and any other variables. The idea is to put every interaction into the model, then delete the item that the associated p-value is greater than 0.05. See figure2 and also see Appendix output4. Lines cross means possibility of interaction. Also, we need to see the p-value from output4. After seeing the result of p-value of each interaction, I will remain terms: gender\*age and gender\*diabetes. P-value is respectively: 0.0088700 and 0.0266578.



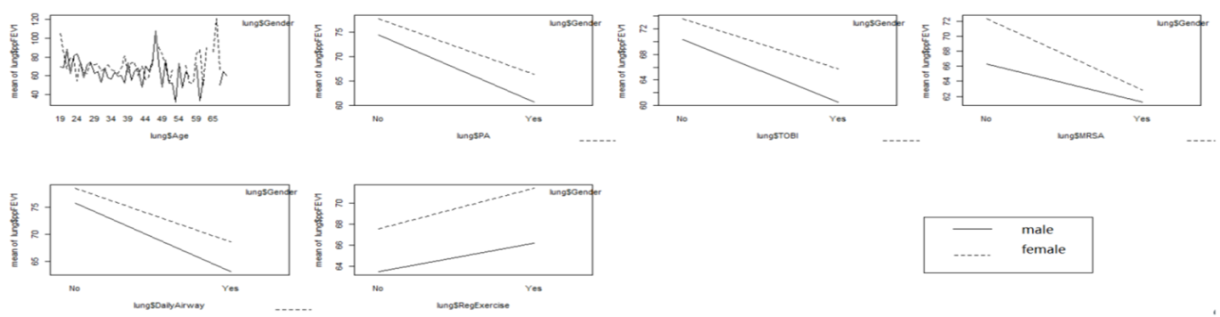


figure 2

We could also separate data into two parts, as male and female groups. So it could fit better, and you may want to analysis male or female separately due to the need of clinical medicine research. See Appendix fit6 and fit7. From summary output, the estimated residuals standard error are respectively: fit3(full data) is 20.13, fit6(only male) is 20.13 and fit7(only female) is 19.77. So it means that if we divide gender by two groups, only for female group, the fit improves a little bit. But it still very slight. If no special need, we won't separate them by two categories.

A second concern is to detect any interactions among age and any other variables. Here, I center the age, it just changes the interpretation of this variable. It replaces age variable by age subtract the mean of age(it is 34). Same idea: put every interactions in the model, then delete the item that the associated p-value is greater than 0.05. See output5 and figure3. So, I will remain terms: gender\*I(age-34) and I(age-34)\*diabetes. P-value is respectively: 0.010847 and 0.035028. See figure3 and Appendix output 5.

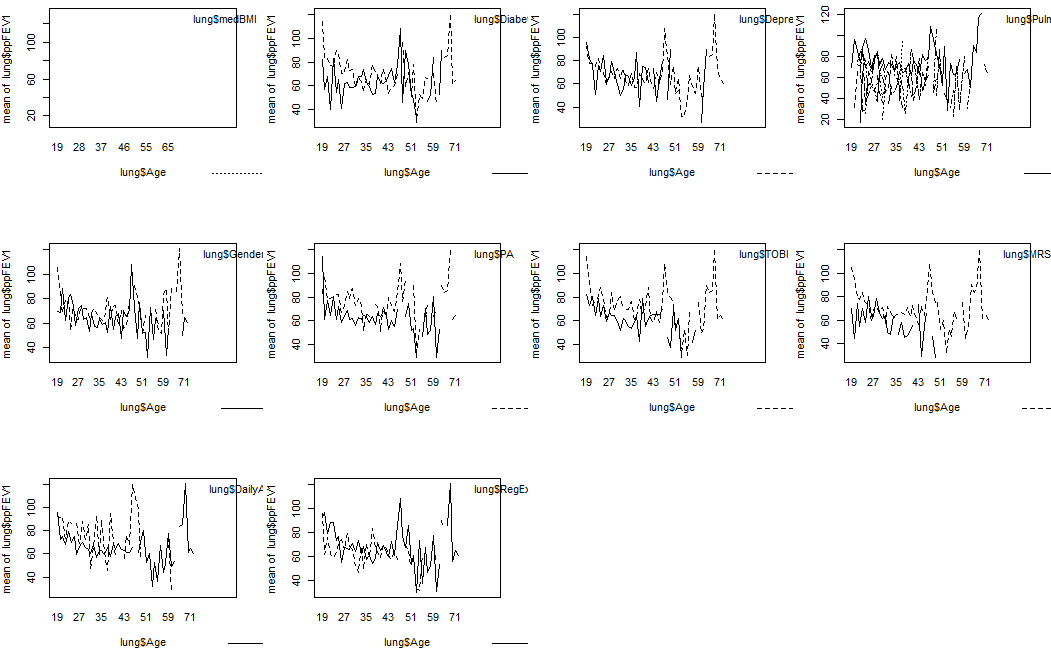


figure 3

Fit9 will include I(age-34)\*gender, I(age-34)\*diabetes, and gender\*diabetes. From ANOVA result between including them and not, p-value is 0.006674.Keep fit9.

Fit10 drop depression if you want a simple model. From ANOVA result between include it and not, p-value is 0.108. Keep fit10.

Fit11 add quadratic form of I(Age-34) into fit10. From ANOVA result between fit10 and

fit11, p-value is 0.002477. Keep fit11.

Besides, I want to test any random effect could be included in the model. Look at the data, a possible idea for random effect is 4 types of races. So, create a new column, and assigned 4 kinds of races by 1 to 4. Then transform them from numeric to factor. Compare two models: one has random effect, the other one does not has. Use gls function for the model does not has random effect. Gls(general least squares) means that allows unequal variance and not independent errors. Use lme function for the model that has random effect. Lme means linear mixed effects model. Both gls and lme use REML(restricted maximum likelihood) by default. While lm function uses ML(maximum likelihood). REML and ML have different algorithm on coefficient estimators. But REML will give unbiased estimates of covariance parameters. See strip chart of races in figure4. From ANOVA result between gls and lme, p-value is 0.4018.

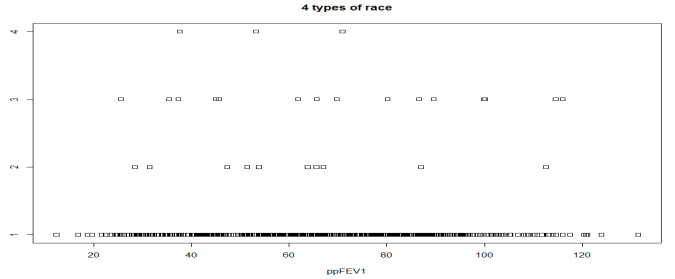


figure 4

So keep fit11 without random effect of race. fit12 is only used to compare with the models which need to detect the estimate of variance or covariance. Because the data of race is very skewed. most of the data is categorized as white. The data is not balanced from 4 types of race and it may not be able to detect the random effect of race. But in the future, if the data could collect more observations besides white race, we could test this issue again.

Now, I want to consider any unequal variance for the different categories within single variable.

First, I concern different variance of gender groups. I use general linear squares function to allow a certain variable has different variance among its levels. Use varIdent command to define it. See fit14 in Appendix part3. From ANOVA result between two models, p-value is 0.9529. Keep fit11 without different variance.

Fit15 includes different variance of diabetes since diabetes has significant interaction terms with both gender and I(age-34). See fit15 in Appendix part3. Also use varIdent command to define it. From ANOVA result between two models(with different variance and without it), p-value is 0.5419. Keep fit11 without different variance.

Fit16 include dependent error into Age. Use varPower command to define it. It means the errors will change due to the levels of that variable, which represents a power variance function structure. See fit16 in Appendix part3. From ANOVA result between two models, p-value is 0.3067. Keep fit11 without different variance.

The final step is to choose a model as simple as possible. When I see the ANOVA and summary of fit11, I decide to test several models that delete some variables which p-value of coefficient is not small than 0.05. Fit18 drops diabetes, p-value is 0.04396. Fit19 deletes MRSA, p-value is 0.1062. Fit20 deletes MARA and diabetes\*gender, p-value is0.09102. Fit21 deletes MARA, age\*diabetes and diabetes\*gender, p-value is 0.03579.

So, final model is fit20. See Appendix part3.

* Diagnostics

We need to do test the hypotheses of "ordinary least square" in linear regression models. So base on that, the best linear unbiased estimators will get from ordinary least squares method. From figure5 we see plot of QQ normal: errors almost all lay on the diagonal line, just jump out a little bit in two tails. Errors are normality basically. And see the plot of residuals v.s. fitted: residuals do not have an obvious pattern. It has equal variance. Figure6 is the correlation plot among numeric variables. These three variables are very weak correlated. Also, we could see gls fit summary result, it shows correlation table for every variable, all correlation of variables are weak(absolute value< 0.4).If multicollinearity exists, it will hurt the accuracy of estimators from "Ordinary linear squares" method. See Appendix output6.

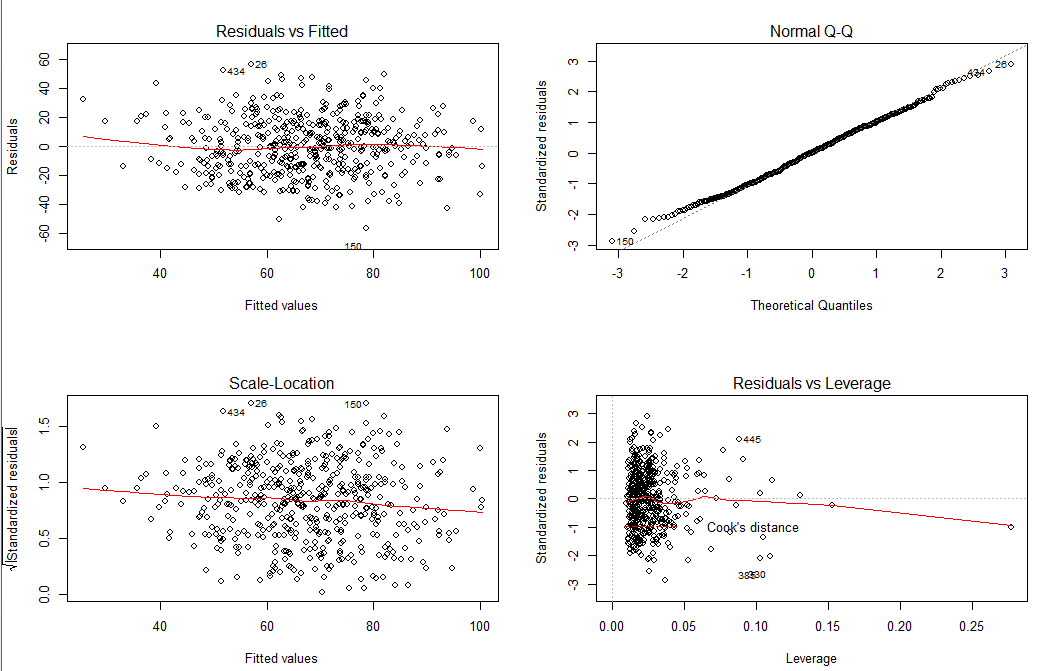
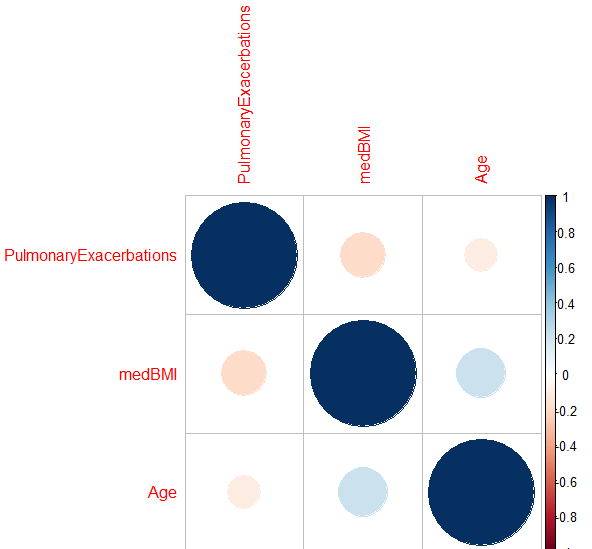
 

figure 5 figure 6

Then, test outliers. Outliers in data can distort predictions and affect the accuracy4. See figure7 for outliers. From the outlier plot, I will choose row:385,330,408,445. Delete them and fit a new model, fit17.

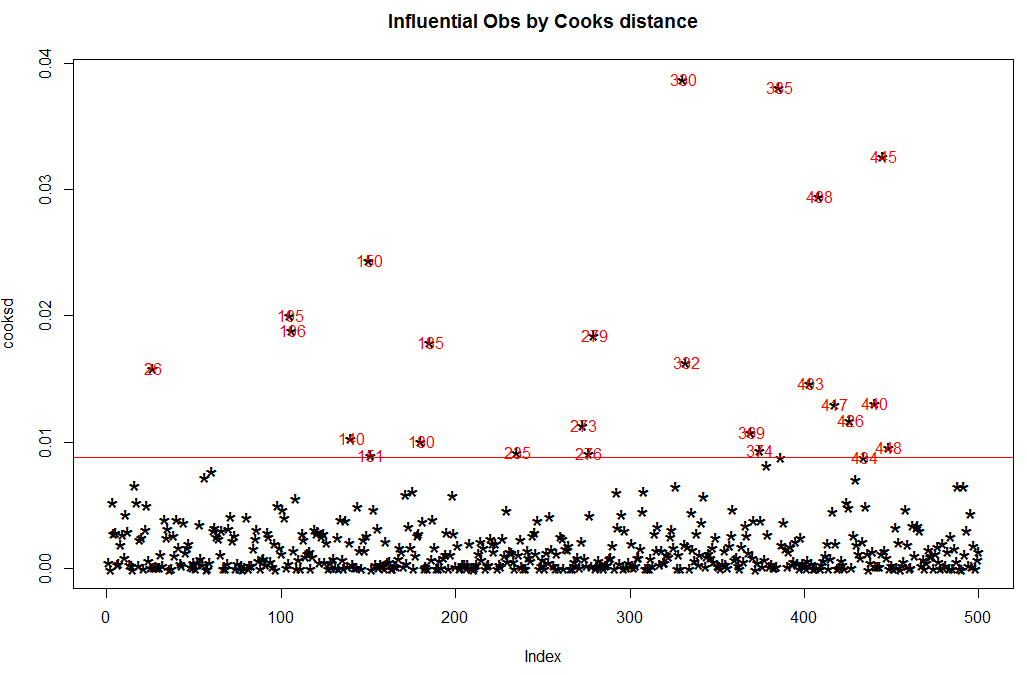
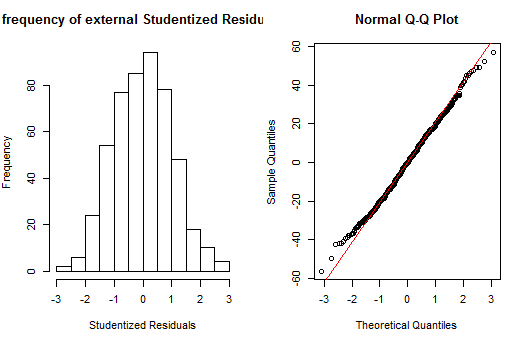
 

figure 7 figure 8

You could do imputation with mean or median4 if you could find which variable has extreme value. But for this data I prefer just delete them and treat them as missing values. Because it is not easy to find which variable affected model appropriately. From summary of two models, residual standard error reduces from 19.87 to 19.65. See Appendix output7. And the diagnostic plot does not show big difference. So keep fit20.

## There are other methods to test three error hypotheses. We could see the plot of external studentized residuals. If it is not bell curve, then residuals are not normality. External studentized residuals mean the algorithm will remove observation i item each time then calculate the residuals. See figure8, residuals are basically normality. I also use Durbin-Watson test in R to test the independence of errors. P-value is 0.4155, we accept null hypothesis and conclude that the errors are independent. And use ncvTest( score test for non-constant error variance), P-value is 0.7527874. we accept null hypothesis and conclude that the errors variance are equal. See Appendix output8.

* Conclusion

See summary of fit20 in Appendix output9. PulmonaryExacerbations(-3.29), I(Age - 34)(-0.58), GenderMale(-6.85), PAYes(-5.58), TOBIYes(-4.47), DailyAirwayYes(-9.17), DiabetesYes(-3.11), I(Age - 34)\*GenderMale(-0.43) have a negative relation with ppFEV1. The associated coefficient is inside parentheses, rounded by 2 digits. It means when those variables value increase one unit the expiratory volume will decrease the amount that is equal to the related coefficient value. DailyAirway and TOBI should be therapies. But why the coefficient is negative, I guess: only the patient who have a very severe problem of breath will need to use an extra therapy to help them feel better from disease symptoms, while the patient who do not have this issue do not need it. When worsening times of infections increase, when patients have PA bacteria and when they have diabetes, their expiratory volume will all have a trend to reduce. Meanwhile, If patients are male and they are old, they have a trend to perform worse than female and young people. Also, the interaction between Age and genderMale shows when male get older, they will perform worse than female. On the other hand, MedBMI(0.79), I((Age - 34)^2)(0.02), RegExerciseYes(5.12) and I(Age - 34)\*DiabetesYes(0.33) have positive relations with ppFEV1. It means daily exercise will increase the expiratory volume.

* Limitation and future studies

From this dataset, the response variable "ppFEV1" only shows us a percentage of normal. We have no clue to know what is the normal value it is based on. Besides, the "ppFEV1" is measured as how much air a person can exhale during a forced breath in the first second. But this method is kind of subjective. It depends on how well the patient did in this process. If one patient did not prepare well and then did not do a full deep breath, or the strength patients used to make a breath are different, the collected results could not be represented properly. Also, people could do a survey to find which therapy is most effective to heal airway infections. The data here is not designed to figure out it.

* Reference

1. The Ohio State University wexner medical center:

https://wexnermedical.osu.edu/lung-pulmonary/ohio-states-lung-center/chronic-pulmonary-infections

2.National Jewish Health:

https://www.nationaljewish.org/conditions/chronic-respiratory-infections

3.wikipedia:

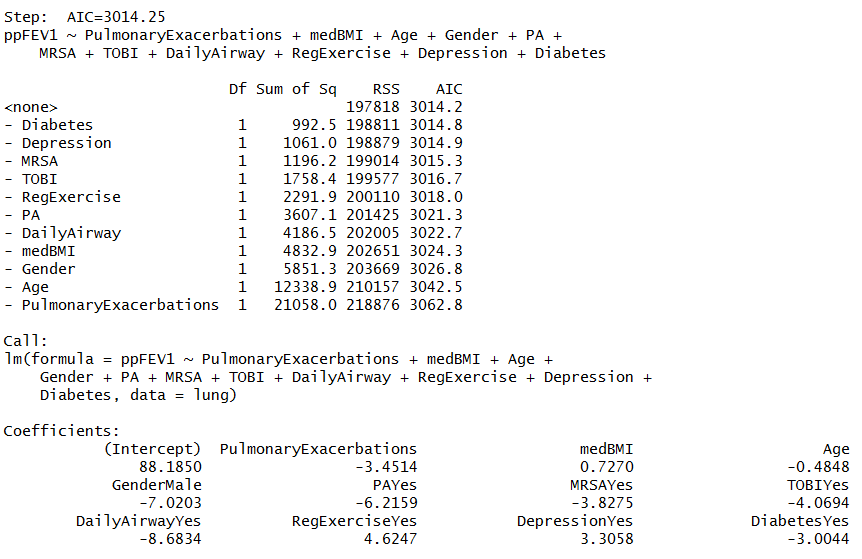
https://en.wikipedia.org/wiki/Multicollinearity

4.statistics.co by Selva Prabahakaran---outliers

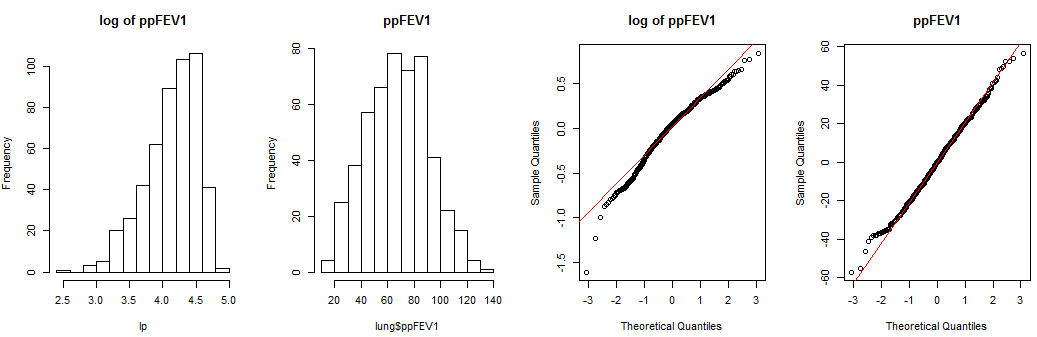
http://r-statistics.co/Outlier-Treatment-With-R.html

* Appendix

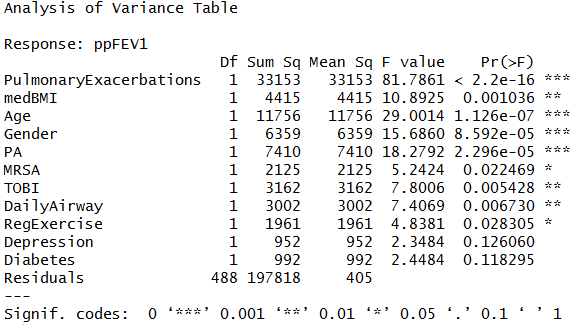
1. Output



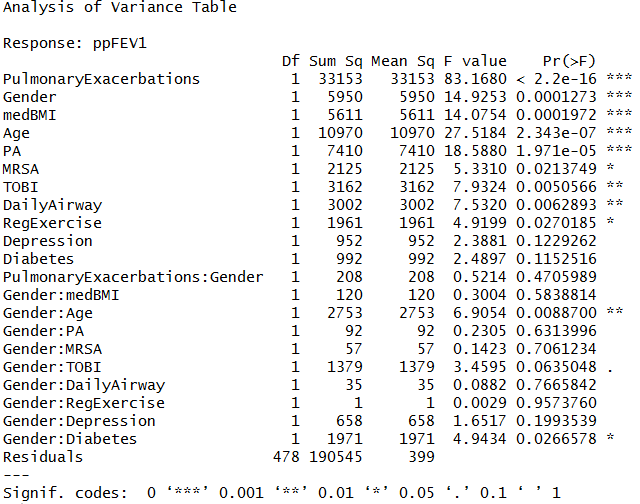
output 1



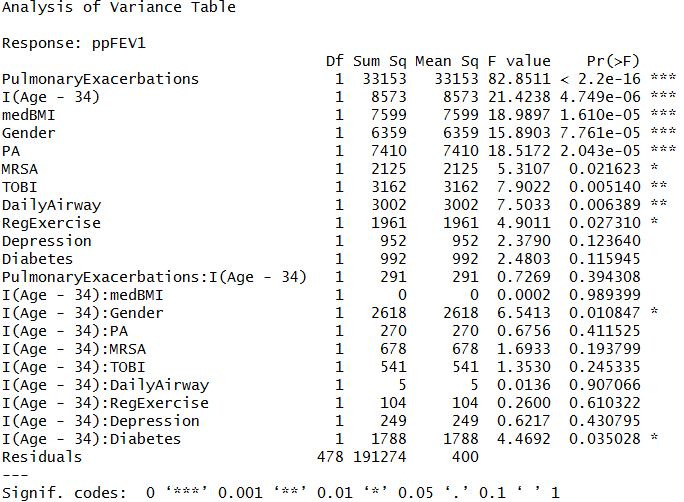
output 2



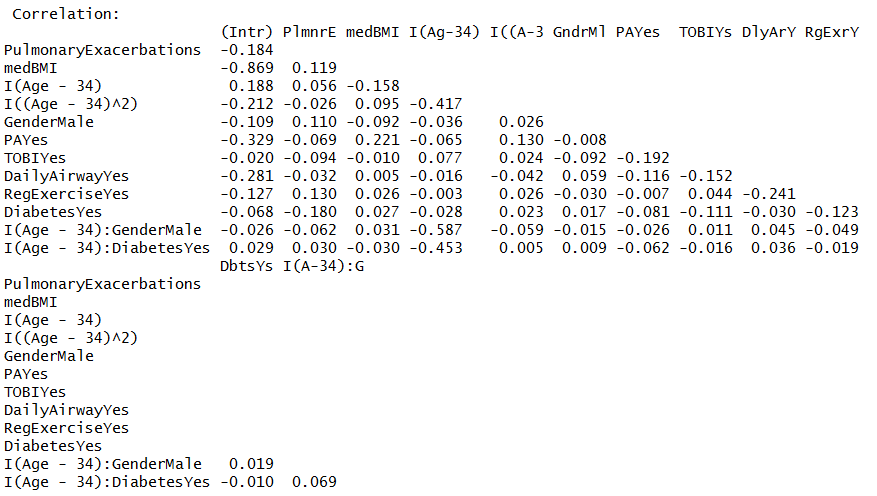
output 3



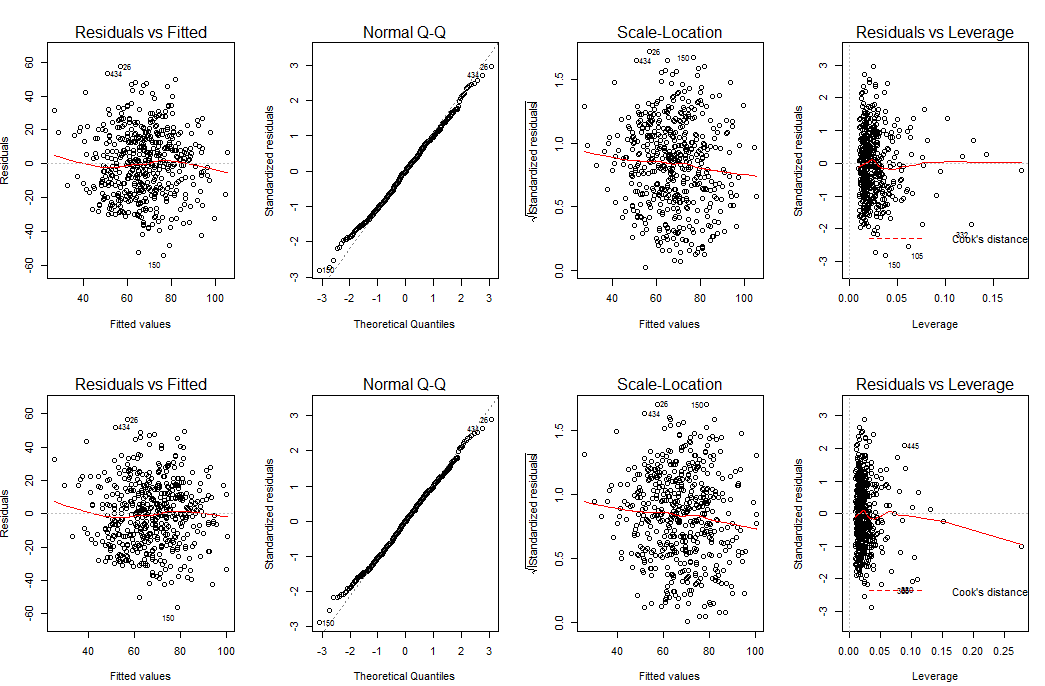
output4



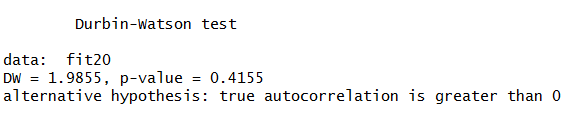
output 5



output 6

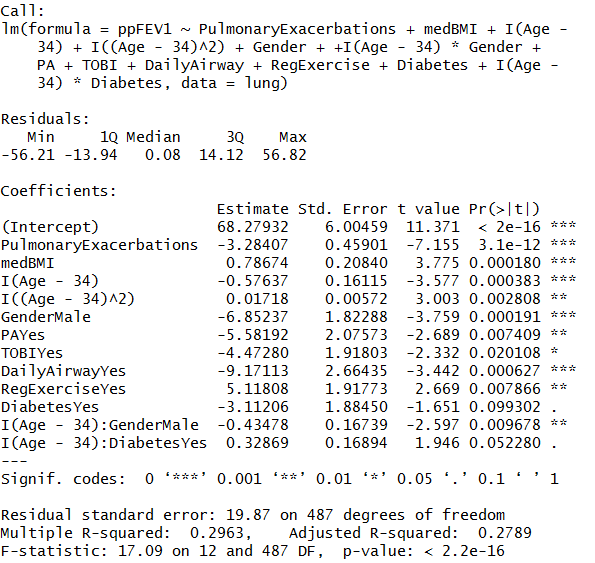


output 7



19.png

output 8



output 9

1. Variables explanations:

LIST OF VARIABLES (in the order of the data columns):

1. ppFEV1: FEV1 percent predicted. FEV is forced expiratory volume and measures how much air a person can exhale during a forced breath. FEV1 is the amount of air exhaled measured during the first second. FEV1 percent predicted is FEV1 converted to a percentage of normal.

2. PulmonaryExacerbations: Number of episodes of acute worsening of airway infections.

3. medHT: median height (centimeters)

4. medWT: median weight (kilograms)

5. medBMI: body mass index calculated using the median height and median weight variables (kg/m2 )

6. Age: age (years)

7. Gender: Male or Female

8. RaceWhite: Patient identifies with racial category White. Yes or No.

9. RaceHispanic: Patient identifies with racial category Hispanic. Yes or No.

10. RaceAA: Patient identifies with racial category African American. Yes or No.

11. RaceAsian: Patient identifies with racial category Asian. Yes or No.

12. DigEnzymes: Nutritional and digestive enzymes are sufficient. Yes or No.

13. PA: Presence of Pseudomonas bacteria. Yes or No.

14. MRSA: Presence of Methicillin Resistant Staphylococcus aureus bacteria. Yes or No.

15. MSSA: Presence of Methicillin Sensitive Staphylococcus aureus bacteria. Yes or No.

16. EColi: Presence of Escherichia coli bacteria. Yes or No.

17. Aspergillus: Presence of Aspergillus fungus. Yes or No.

18. HypertonicSaline: Use of hypertonic saline mist therapy. Yes or No.

19. TOBI: Use of inhaled tobramycin antibiotic. Yes or No.

20. DailyAirway: Use of daily airway clearance techniques. Yes or No.

21. RegExercise: Participates in daily exercise. Yes or No.

22. Depression: Diagnosed with depression. Yes or No.

23. Diabetes: Diagnosed with diabetes. Yes or No

1. Code:

lung<-read.csv(file="DataFile\_Sp18.csv",header=TRUE)

attach(lung)

library(nlme)

library(corrplot)

library(car)

install.packages("lmtest")

library(lmtest)

library(MASS)

summary(fit1)

step(fit1,trace=TRUE)

range(medBMI)

range(Age)

range(PulmonaryExacerbations)

fit1<-lm(ppFEV1~.,data=lung)

par(mfrow=c(3,4))

hist(lung$ppFEV1,main="ppFEV1")

plot(as.factor(lung$PulmonaryExacerbations),

main="times of worsening infections")

hist(lung$medBMI,main="BMI")

hist(lung$Age,main="age")

plot(lung$Gender,main="gender")

plot(lung$PA,main="pseudomonas bacteria")

plot(lung$MRSA,main="MRSA bacteria")

plot(lung$TOBI,main="inhaled tobramycin antibiotic")

plot(lung$DailyAirway,main="daily airway clearance techniques")

plot(lung$RegExercise,main="daily exercise")

plot(lung$Depression,main="depression")

plot(lung$Diabetes,main="diabetes")

par(mfrow=c(2,2))

lp<-log(lung$ppFEV1)

hist(lp)

hist(lung$ppFEV1)

fit2<-lm(lp~PulmonaryExacerbations+medBMI+Age+Gender

+PA+MRSA+TOBI+DailyAirway+RegExercise+Depression+Diabetes,data=lung)

summary(fit2)

length(which(lung$PulmonaryExacerbations==1))

fit3<-lm(ppFEV1~PulmonaryExacerbations+medBMI+Age+Gende

+PA+MRSA+TOBI+DailyAirway+RegExercise+Depression+Diabetes,data=lung)

summary(fit3)

ANOVA(fit3)

AIC(fit2,fit3)

# no use log transformation, due to violation of normality hypothesis of residuals

# in linear models. So even the AIC of log fit is way more smaller

# than the original form, I won't choose it.

qqnorm(fit2$residuals)

qqline(fit2$residuals,col="red")

qqnorm(fit3$residuals)

qqline(fit3$residuals,col="red")

par(mfrow=c(2,2))

plot(fit2)

plot(fit3)

ANOVA(fit3)

#fit4 drop depression and diabetes.

fit4<-lm(ppFEV1~PulmonaryExacerbations+medBMI+Age+Gender+PA+MRSA+TOBI

+DailyAirway+RegExercise,data=lung)

AIC(fit3,fit4)

ANOVA(fit3,fit4)

# compare two models, since p-value are all greater than 0.05,

# the results show that they are not significant different,

# if you want to have a simple model, then use fit4.

# if you want to follow the rule to choose a smaller AIC, then use fit3.

plot(fit4)

plot(fit3$fitted.values,ppFEV1)

lines(x=c(0,100),y=c(0,100),col="red")

# test any interaction terms exists

# interactions among gender and any other variables

# put every interactions in the model, then delete the item

# that the associated p-value is greater than 0.05.

interaction.plot(lung$medBMI,lung$Gender,lung$ppFEV1)

interaction.plot(lung$Diabetes,lung$Gender,lung$ppFEV1)

interaction.plot(lung$Depression,lung$Gender,lung$ppFEV1)

interaction.plot(lung$PulmonaryExacerbations,lung$Gender,lung$ppFEV1)

interaction.plot(lung$Age,lung$Gender,lung$ppFEV1)

interaction.plot(lung$PA,lung$Gender,lung$ppFEV1)

interaction.plot(lung$TOBI,lung$Gender,lung$ppFEV1)

interaction.plot(lung$MRSA,lung$Gender,lung$ppFEV1)

interaction.plot(lung$DailyAirway,lung$Gender,lung$ppFEV1)

interaction.plot(lung$RegExercise,lung$Gender,lung$ppFEV1)

par(mfrow=c(1,1))

fit5<-lm(ppFEV1~PulmonaryExacerbations+PulmonaryExacerbations\*Gender

+medBMI+medBMI\*Gender+Age+Age\*Gender+Gender+PA+PA\*Gender

+MRSA+MRSA\*Gender+TOBI+TOBI\*Gender+DailyAirway+DailyAirway\*Gender

+RegExercise+RegExercise\*Gender+Depression+Depression\*Gender+Diabetes

+Gender\*Diabetes,data=lung)

ANOVA(fit5)

# only remain:gender\*age,gender\*diabetes.

#separate data by two parts, as male and female groups

male<- lung[which(lung$Gender=="Male"),]

female<-lung[which(lung$Gender=="Female"),]

fit6<-lm(ppFEV1~PulmonaryExacerbations+medBMI+Age

+PA+MRSA+TOBI+DailyAirway+RegExercise+Depression+Diabetes,data=male)

fit7<-lm(ppFEV1~PulmonaryExacerbations+medBMI+Age

+PA+MRSA+TOBI+DailyAirway+RegExercise+Depression+Diabetes,data=female)

AIC(fit3,fit6,fit7)

summary(fit6)

summary(fit7)

summary(fit3)

# center the age, it just changes the interpretation of this variable.

# test any interaction terms exists

# interactions among age and any other variables

# put every interactions in the model, then delete the item

# that the associated p-value is greater than 0.05.

interaction.plot(lung$Age,lung$medBMI,lung$ppFEV1)

interaction.plot(lung$Age,lung$Diabetes,lung$ppFEV1)

interaction.plot(lung$Age,lung$Depression,lung$ppFEV1)

interaction.plot(lung$Age,lung$PulmonaryExacerbations,lung$ppFEV1)

interaction.plot(lung$Age,lung$Gender,lung$ppFEV1)

interaction.plot(lung$Age,lung$PA,lung$ppFEV1)

interaction.plot(lung$Age,lung$TOBI,lung$ppFEV1)

interaction.plot(lung$Age,lung$MRSA,lung$ppFEV1)

interaction.plot(lung$Age,lung$DailyAirway,lung$ppFEV1)

interaction.plot(lung$Age,lung$RegExercise,lung$ppFEV1)

mean(Age)

fit8<-lm(ppFEV1~PulmonaryExacerbations+PulmonaryExacerbations\*I(Age-34)

+medBMI+medBMI\*I(Age-34)+I(Age-34)+Gender+I(Age-34)\*Gender

+PA+PA\*I(Age-34)+MRSA+MRSA\*I(Age-34)+TOBI+TOBI\*I(Age-34)

+DailyAirway+DailyAirway\*I(Age-34)+RegExercise+RegExercise\*I(Age-34)

+Depression+Depression\*I(Age-34)+Diabetes+Diabetes\*I(Age-34),data=lung)

ANOVA(fit8)

#only remain:age\*gender,age\*diabetes.

#So fit9 will include age\*gender,age\*diabetes,gender\*diabetes

fit9<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+Gender

+PA+MRSA+TOBI+DailyAirway+RegExercise+Depression+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,data=lung)

ANOVA(fit3,fit9)

AIC(fit3,fit9)

# fit9 is better

#fit10 drops depression if you want a simple model

fit10<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+Gender

+PA+MRSA+TOBI+DailyAirway+RegExercise+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,data=lung)

ANOVA(fit9,fit10)

AIC(fit9,fit10)

#keep fit10

# fit11 add quadratic form of I(Age-34) into fit10.

fit11<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+MRSA+TOBI+DailyAirway+RegExercise+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,data=lung)

ANOVA(fit11,fit10)

AIC(fit11,fit10)

# keep fit 11

# try to test any random effect exist

# look at the data, the most possible idea for random effect

# is the 4 types of race.

# So, create a new column, and assigned by 1 to 4.

# that represent 4 kinds of races.

lung$Race[which(lung$RaceWhite=="Yes")]<-1

lung$Race[which(lung$RaceHispanic=="Yes")]<-2

lung$Race[which(lung$RaceAA=="Yes")]<-3

lung$Race[which(lung$RaceAsian=="Yes")]<-4

lung$Race<-as.factor(lung$Race)

stripchart(ppFEV1~Race,data=lung,main="4 types of race")

# compare two models: one has random effect, the other one do not has.

# gls means general least squares that allows unequal variance and not independent errors.

fit12<-gls(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+MRSA+TOBI+DailyAirway+RegExercise+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,data=lung)

fit13<-lme(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+MRSA+TOBI+DailyAirway+RegExercise

+Diabetes+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,random=~1|Race,data=lung)

ANOVA(fit12,fit13)

# So keep fit 11 without random effect of race.

# Because the data of race is very skewed. most of the data is categorized as white.

# the data is not balanced from 4 types of race and may not be able

# to detect the random effect of race.

# but in the future, if the data could collect more observations besides white race

# we could test this again.

#test different variance of gender groups.

fit14<-gls(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+MRSA+TOBI+DailyAirway+RegExercise+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,

weights=varIdent(form=~1|Gender),data=lung)

ANOVA(fit12,fit14)

# keep fit 11 without different variance in gender groups.

# fit15 include different variance of diabetes, since diabetes has significant

# interaction terms with both gender and I(age-34).

fit15<-gls(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+MRSA+TOBI+DailyAirway+RegExercise+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,

weights=varIdent(form=~1|Diabetes),data=lung)

ANOVA(fit12,fit15)

#keep fit 11 without difference variance in diabetes groups.

# fit16 include dependent error into Age

# varPower means the errors will change due to the levels of that variable.

# representing a power variance function structure.

fit16<-gls(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+MRSA+TOBI+DailyAirway+RegExercise+Diabetes

+I(Age-34)\*Gender+I(Age-34)\*Diabetes+Gender\*Diabetes,

weights=varPower(form=~Age),data=lung)

ANOVA(fit12,fit16)

# keep fit11.

ANOVA(fit11)

summary(fit11)

#from avona of fit11

# fit18 drop diabetes from fit11 . if you want to get a simple model.

fit18<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+I(Age-34)\*Gender+PA+MRSA+TOBI+DailyAirway+RegExercise,data=lung)

ANOVA(fit18,fit11)

#keep fit11

#from summary of fit11

#fit21 delete MARA, Age\*diabetes and diabetes\*gender

fit21<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+I(Age-34)\*Gender+PA+TOBI+DailyAirway+RegExercise

+Diabetes,data=lung)

ANOVA(fit21,fit11)

# keep fit11

#fit19 delete MRSA

fit19<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+I(Age-34)\*Gender+PA+TOBI+DailyAirway+RegExercise

+Diabetes+I(Age-34)\*Diabetes+Gender\*Diabetes,data=lung)

ANOVA(fit19,fit11)

#keep fit19

#fit20 delete MARA and diabetes\*gender

fit20<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+I(Age-34)\*Gender+PA+TOBI+DailyAirway+RegExercise

+Diabetes+I(Age-34)\*Diabetes,data=lung)

ANOVA(fit20,fit11)

summary(fit20)

#keep fit20

# some types of test to test the hypotheses of "ordinary least square"

# in linear regression models

# So base on that, the best linear unbiased estimators will get from

# ordinary least squares method

par(mfrow=c(2,2))

plot(fit20)

# detect multicollinearity.

# if multicollinearity exist, it will hurt the accuracy of estimators

# from "Ordinary linear squares" method

# see correlation for gls

fit22<-gls(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+I(Age-34)\*Gender+PA+TOBI+DailyAirway+RegExercise

+Diabetes+I(Age-34)\*Diabetes,data=lung)

summary(fit22)

#variables need to be numerical values in the correlation test(cor or corrplot).

#then test numerical variables' correlation

#results show that they are reasonably independent.

corrplot(cor(data.frame(PulmonaryExacerbations,medBMI,Age)))

cor(data.frame(PulmonaryExacerbations,medBMI,Age))

#Outliers in data can distort predictions and affect the accuracy

outlierTest(fit20)

# Most extreme observation is row 26.

#In general use, those observations that

#have a cook’s distance greater than 4 times the mean may be classified as influential.

cooksd<-cooks.distance(fit20)

plot(cooksd, pch="\*", cex=2, main="Influential Obs by Cooks distance")

# plot cook's distance

abline(h = 4\*mean(cooksd, na.rm=T), col="red")

# add cutoff line

text(x=1:length(cooksd)+1, y=cooksd,

labels=ifelse(cooksd>4\*mean(cooksd, na.rm=T),names(cooksd),""), col="red")

# add labels

# you could do imputation with mean / median / mode if you could find which variable

# has extreme value and then impute with mean or median.

# But for this data I prefer just delete them and treat them as missing values.

# because for the extreme case you cannot find

# the extreme variable that exactly affected model.

# here, from the outlier plot,

# I will choose row:385,330,408,445. delete them and fit a new model fit17.

newlung<-lung[-c(385,330,408,445),]

fit17<-lm(ppFEV1~PulmonaryExacerbations+medBMI+I(Age-34)+I((Age-34)^2)

+Gender+PA+TOBI+DailyAirway+RegExercise+I(Age-34)\*Gender

+I(Age-34)\*Diabetes,data=newlung)

plot(fit17)

plot(fit20)

summary(fit17)

summary(fit20)

# diagnostic results do not show big difference.

# keep fit20 that include all observations.

# first hypothesis: errors are normal distribution.

# external means remove observation i item then calculate the residuals.

# plot shows it basically meets the requirement of normality.

hist(studres(fit20), main="frequency of external Studentized Residuals",

xlab="Studentized Residuals")

qqnorm(fit20$residuals)

qqline(fit20$residuals,col="red")

# second hypothesis: errors are independent.

dwtest(fit20)

#because p-value is 0.4155,

#we accept null hypothesis, we conclude that the errors are independent.

# third hypothesis: errors have equal variance.

ncvTest(fit20)

#because p-value is 0.7527874,

#we accept null hypothesis, we conclude that the errors' variance is equal.

##### So, final model is fit20.#####