36401-B Data Analysis Exam 3

Joong Ho Choi

TOTAL POINTS

100 / 100

QUESTION 1

1 Introduction 20 / 20

✓ **- 0 pts** **Correct**

- 20 pts Missing

QUESTION 2

2 EDA 25 / 25

✓ **- 0 pts** **Correct**

- 4 pts Needs further investigation of univariate relationships (summary statistics, analysis of missing values etc.)

- 4 pts Needs further investigation of bivariate relationships such as scatterplots between independent and dependent variables

- 4 pts Hard to see you analysis visually -- figures / tables not legible, or scatterplots are inappropriate for type of data

- 4 pts Lacks sufficient analysis/commentary on plots.

- 8 pts Needs substantial additions - ie. more analysis of what the graphs actually tell us.

- 4 pts Modelling should come after EDA

1 Cool graph - what do the colors represent?

QUESTION 3

3 Modeling 30 / 30

✓ **- 0 pts** **Correct**

- 0 pts When fitting parametric models, it would be better to mention a little bit why you didn't use higher-order terms, interaction terms, transformation, etc. Say, from EDA we don't see any particular nonlinear relationship, or I tried a few transformations but they didn't help. (For nonparametric models, no need to try transformation, etc because they would essentially transform the variables.)

- 2 pts Didn't calculate/show the confidence interval of causal effect.

- 3 pts When doing causal inference, it's important to include all possible confounding variables as the covariates, even if those variables may not be significant/do not have a clear relationship with the outcome in EDA. We may estimate the causal effect more precisely, in this way.

- 1 pts It would be better to use a test set to evaluate models that risk overfitting the data. In this dataset, training error can be very different from test error.

- 2 pts A linear model is not appropriate for binary outcome.

- 2 pts Didn't fit the counts when using the log-linear model.

- 1 pts "Most of the variables are binary" seems incorrect to me. Only three predictors, sex, pg, and drug, are binary.

- 1 pts Semiparametric models mostly refer to models that have both parametric parts and non-parametric parts, like additive models, etc. Random forests should be non-parametric models.

- 2 pts Accuracy shouldn't be below 0.5

- 1 pts I think you meant 'error' when you said "Accuracy".

- 2 pts From the first table on page 6, errors of logistic model, tree and random forest are 0.30,0.35, 0.28. Why did you say logistic model has the smallest error?

- 1 pts Prediction error shouldn't be larger than 50% if you meant the fraction of misclassified data points.

- 1 pts If the outcome is binary, linear models normally wouldn't be better than logistic models. If you are pretty sure your results are correct, please ask for a regrade.

- **3 pts** Error with the plugging-in estimator. The two estimates shouldn't be the same.

- **2 pts** Log-linear model is used to find the causal relationships, not to predict the outcome quality. Comparing log-linear models with logistic and additive models doesn't make sense to me.

- **5 pts** Didn't use quantitative analysis (like plug-in estimator) to estimate the causal effect. Correlations are not causality

- **15 pts** Sorry I didn't find your mode for goal 2 and goal 3

QUESTION 4

4 Final Models **25 / 25**

✓ **- 0 pts Correct**

- **2 pts** Conclusions should contain a more elaborate discussion of the results of the final models, trying to connect them to the actual data and the problem you were required to solve. For instance, for the causal inference part you could have told more about why you think your model did or did not uncover any causal effect, commenting on possible omitted-variable bias or on the plausibility in a real-world setting of the effect you found, if any.

- **5 pts** No conclusion about predictive model

- **5 pts** No conclusions about causal inference

- **5 pts** No conclusions about graphical models

- **8 pts** No limitations

- **3 pts** Taking the average of training and test error is wrong. Training error underestimates the true generalization error, hence if you had a model that strongly overfits the data you might end up with a suboptimal choice

- **4 pts** Insufficient explanation in limitations

- **8 pts** No discussion of results, just reported some metrics

- **25 pts** Missing entirely

36401test3

R Markdown

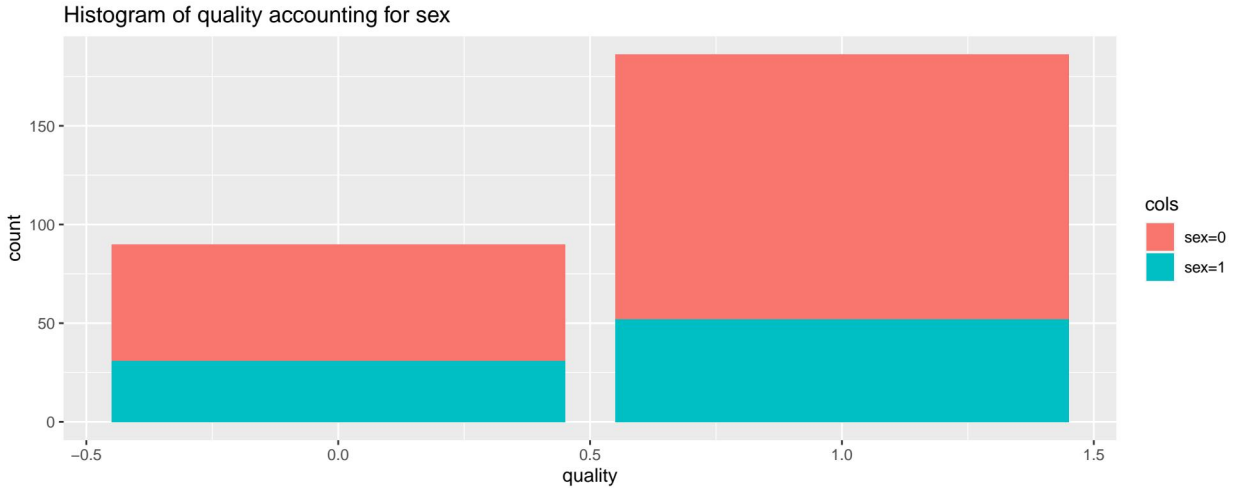
Introduction

```
##   pg age sex educ insu drug severity com     pcsd   mcs.sd quality
## 1  1  36   0    5    1    1        2   4 49.97533 49.57280       1
## 2  1  37   1    6    2    1        3   3 40.93979 56.11393       1
## 3  1  43   1    5    1    1        4   3 50.32495 57.11961       1
```

The data comes from a file called asthma.txt, which contains data on 276 patients in California.The outcome of interest is quality. This is a binary variable indicating whether the patient was satisfied with their treatment (1= satsified, 0 = not satsified). Excluding quality, the data has the following 10 variables: pg (physician group), age, sex, educ(education), insu (insurance status), drug (drug coverage status), severity, com (comorbidity), pcs(physical comorbidity scale), mcs (mental comorbidity scale).

There are 3 main research problems: (1) predicting quality from the other variables. (2) using causal inference to infer the causal effect of physician group on patient satisfaction. (3) using graphical modeling to understand the relationship between the binary variables(pg,sex and quality)

Exploratory Data Analysis



1 Introduction **20 / 20**

✓ **- 0 pts** **Correct**

- 20 pts Missing

36401test3

R Markdown

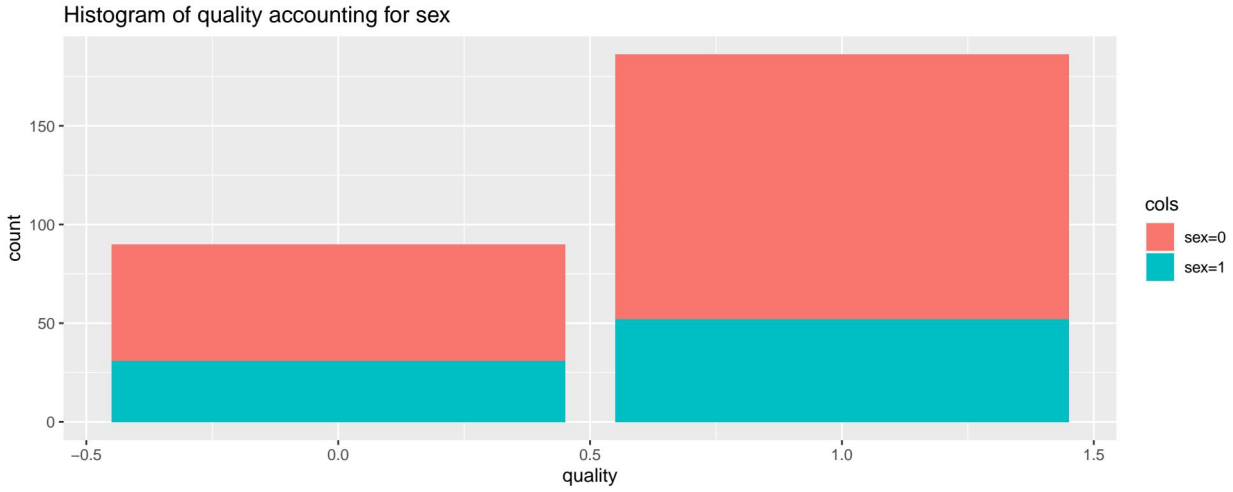
Introduction

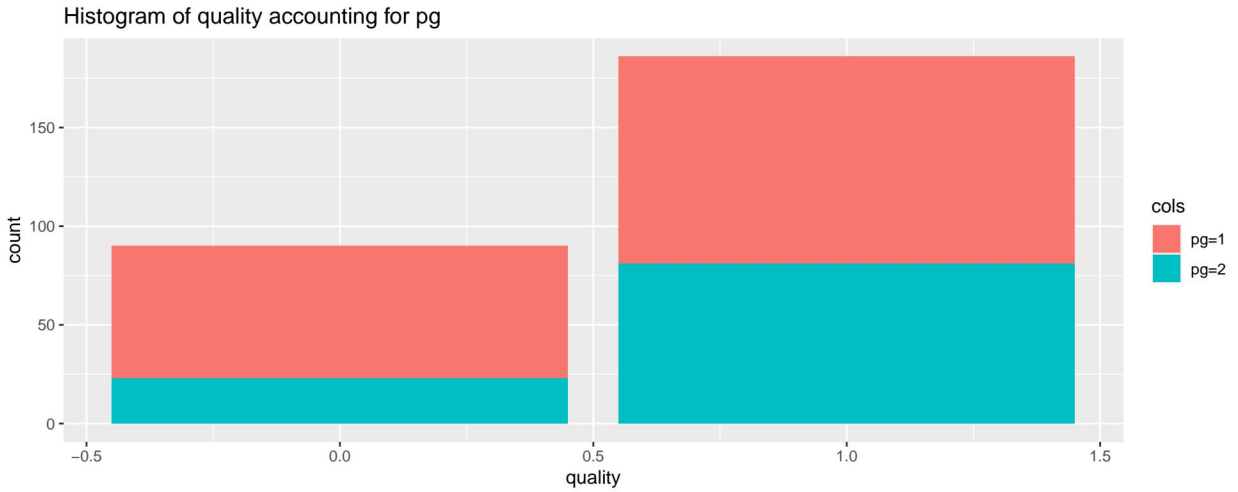
```
##   pg age sex educ insu drug severity com     pcsd   mcs.sd quality
## 1  1  36   0    5    1    1        2   4 49.97533 49.57280       1
## 2  1  37   1    6    2    1        3   3 40.93979 56.11393       1
## 3  1  43   1    5    1    1        4   3 50.32495 57.11961       1
```

The data comes from a file called asthma.txt, which contains data on 276 patients in California.The outcome of interest is quality. This is a binary variable indicating whether the patient was satisfied with their treatment (1= satsified, 0 = not satsified). Excluding quality, the data has the following 10 variables: pg (physician group), age, sex, educ(education), insu (insurance status), drug (drug coverage status), severity, com (comorbidity), pcs(physical comorbidity scale), mcs (mental comorbidity scale).

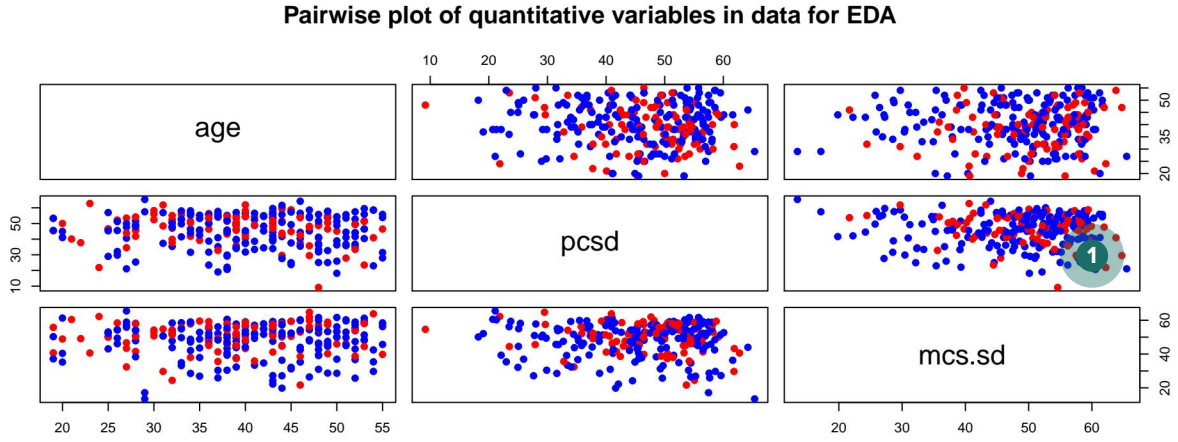
There are 3 main research problems: (1) predicting quality from the other variables. (2) using causal inference to infer the causal effect of physician group on patient satisfaction. (3) using graphical modeling to understand the relationship between the binary variables(pg,sex and quality)

Exploratory Data Analysis





Before performing EDA, we ensured there is no missing value in the data. Then, we inspected quality, pg and sex as they are the variables of interest for research question 3. From the histogram above, we see that the number of people who reported satisfied with treatment is about twice of that of people who reported unsatisfied with the treatment. Also, the count of sex=0 is higher than that of sex=1 for both quality; the count of pg=1 is higher than that of pg=2 for both quality as well.



Then, for research question 1, I decided to inspect the relationship between 3 continuous variables(age,pcsd and mcs.sd), while accounting for two outcomes of quality(the variable of interest); red points represent quality=0 and blue points represent quality=1. However, as you can see from the scatterplots above, there is no clear linear relationship between the variables for both outcomes of quality; it is difficult to distinguish between quality=0 and quality=1 as many red and blue points overlap with each other.

Modeling

For modeling, I transformed discrete variables(educ, insu, drug, severity) into factors in order to help models learn data's pattern more easily; I decided not to transform com, a discrete variable, into factor because it worsened the performance in general; I kept the other variables same. From the EDAs above, it seems the variable 'quality' shares non-linear relationships with continuous variables(age,pcsd,mcs.sd). Since the rest of the variables are integer types that are either binary or of short-ranged, I decided that random forest would be a good model to predict quality from the other variables. To test its performance, I used a train/test split.

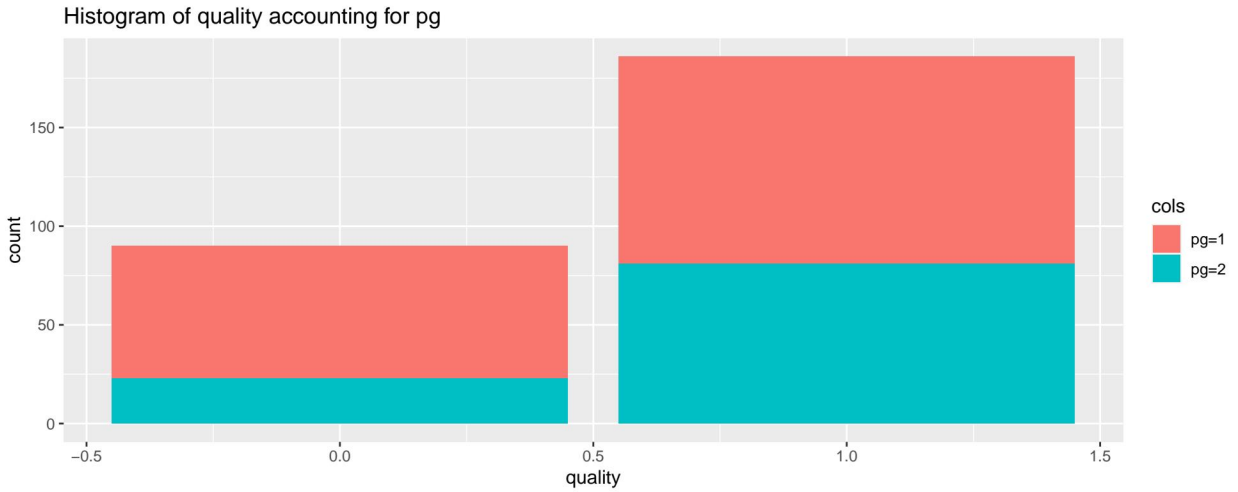
It showed 66.67% accuracy. Then, for comparison, I experimented with logistic classifier and generalised additive model. As additive model already involves smoothing, I did not apply any transformation on the

2 EDA 25 / 25

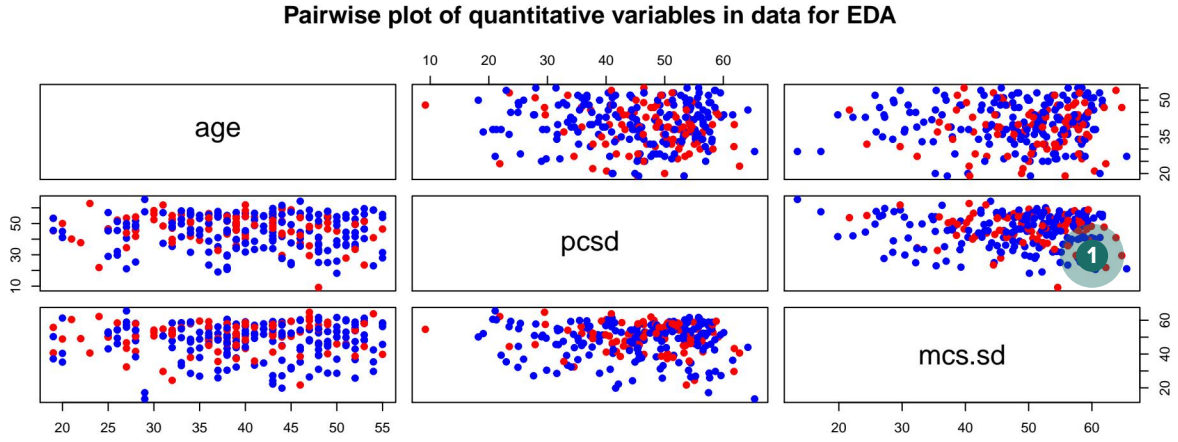
✓ - 0 pts Correct

- 4 pts Needs further investigation of univariate relationships (summary statistics, analysis of missing values etc.)
- 4 pts Needs further investigation of bivariate relationships such as scatterplots between independent and dependent variables
- 4 pts Hard to see you analysis visually -- figures / tables not legible, or scatterplots are inappropriate for type of data
- 4 pts Lacks sufficient analysis/commentary on plots.
- 8 pts Needs substantial additions - ie. more analysis of what the graphs actually tell us.
- 4 pts Modelling should come after EDA

1 Cool graph - what do the colors represent?



Before performing EDA, we ensured there is no missing value in the data. Then, we inspected quality, pg and sex as they are the variables of interest for research question 3. From the histogram above, we see that the number of people who reported satisfied with treatment is about twice of that of people who reported unsatisfied with the treatment. Also, the count of sex=0 is higher than that of sex=1 for both quality; the count of pg=1 is higher than that of pg=2 for both quality as well.



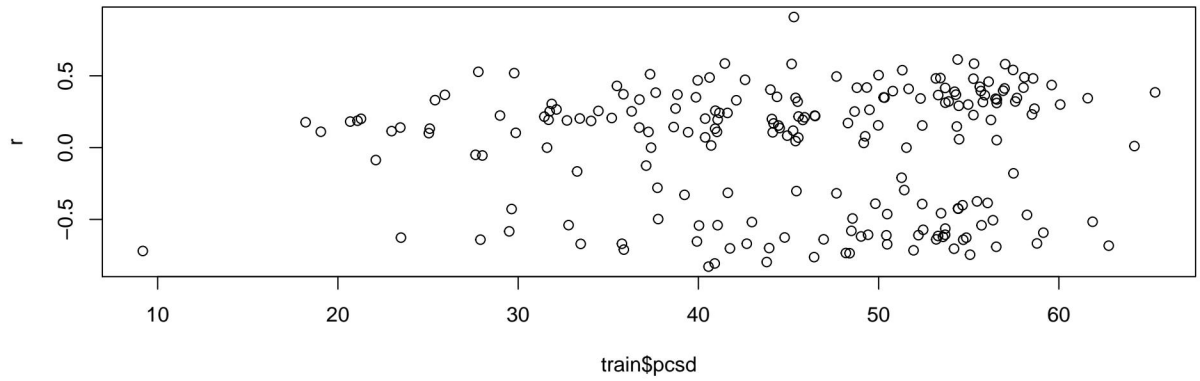
Then, for research question 1, I decided to inspect the relationship between 3 continuous variables(age,pcsd and mcs.sd), while accounting for two outcomes of quality(the variable of interest); red points represent quality=0 and blue points represent quality=1. However, as you can see from the scatterplots above, there is no clear linear relationship between the variables for both outcomes of quality; it is difficult to distinguish between quality=0 and quality=1 as many red and blue points overlap with each other.

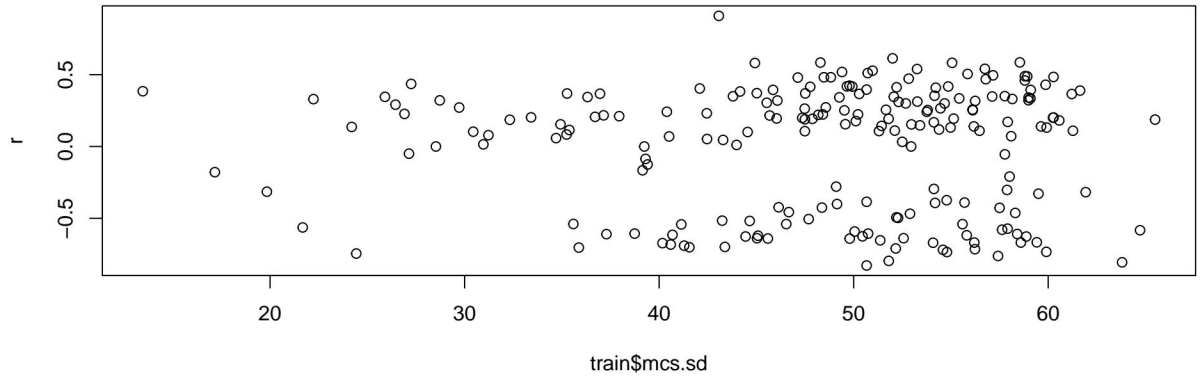
Modeling

For modeling, I transformed discrete variables(educ, insu, drug, severity) into factors in order to help models learn data's pattern more easily; I decided not to transform com, a discrete variable, into factor because it worsened the performance in general; I kept the other variables same. From the EDAs above, it seems the variable 'quality' shares non-linear relationships with continuous variables(age,pcsd,mcs.sd). Since the rest of the variables are integer types that are either binary or of short-ranged, I decided that random forest would be a good model to predict quality from the other variables. To test its performance, I used a train/test split.

It showed 66.67% accuracy. Then, for comparison, I experimented with logistic classifier and generalised additive model. As additive model already involves smoothing, I did not apply any transformation on the

variables. Since quality,pg,sex,educ,insu,drug,severity,com are discrete variables, no smoothing was applied. Overall, log model and additive model achieved 62.32% and 65.22% accuracy respectively.

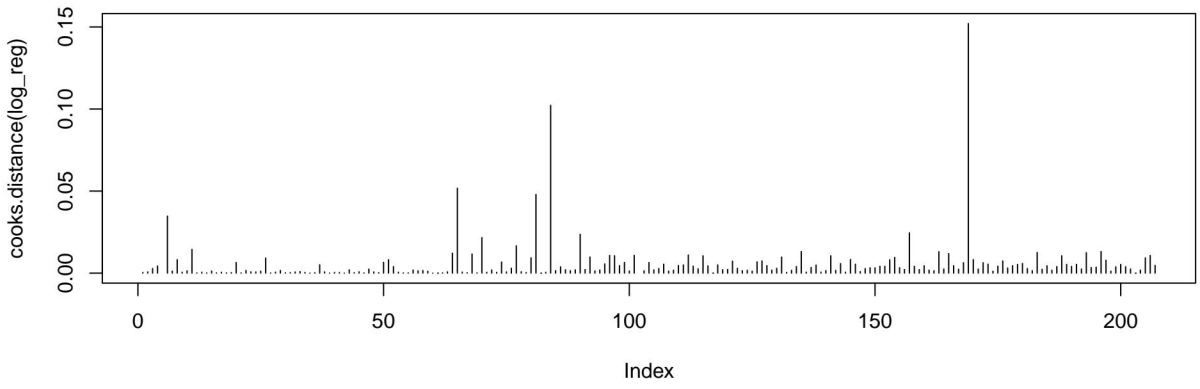


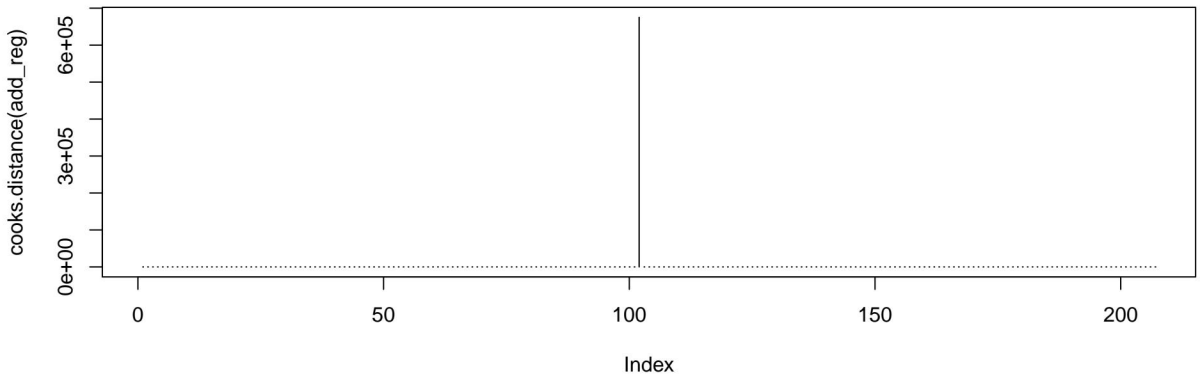



```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## quality ~ pg + s(age) + sex + educ + insu + drug + severity +
##     com + s(pcsd) + s(mcs.sd)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0807696  0.5263012   0.153  0.87820
## pg          -0.2298722  0.0775949  -2.962  0.00345 **
## sex          0.0009129  0.0735201   0.012  0.99011
## educ2        1.0409050  0.5245018   1.985  0.04865 *
## educ3        0.8738024  0.4829854   1.809  0.07203 .
## educ4        0.7525902  0.4697328   1.602  0.11080
## educ5        0.8465980  0.4690599   1.805  0.07270 .
## educ6        0.7407423  0.4691561   1.579  0.11605
## insu2        0.0906567  0.0749240   1.210  0.22781
```

```
## insu3        0.1564163  0.4714763   0.332  0.74044
## insu5       -0.0944827  0.1639290  -0.576  0.56506
## drug1        0.3383546  0.2177143   1.554  0.12184
## severity2   -0.1370856  0.1108098  -1.237  0.21759
## severity3   -0.2221633  0.0975780  -2.277  0.02393 *
## severity4   -0.2950250  0.1277571  -2.309  0.02202 *
## com          0.0024116  0.0260038   0.093  0.92621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(age)    1.612  2.008 0.838  0.4379
## s(pcsd)   1.000  1.000 1.861  0.1742
## s(mcs.sd) 1.000  1.000 4.252  0.0406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0892   Deviance explained = 17.1%
## GCV = 0.22133  Scale est. = 0.20036   n = 207
```

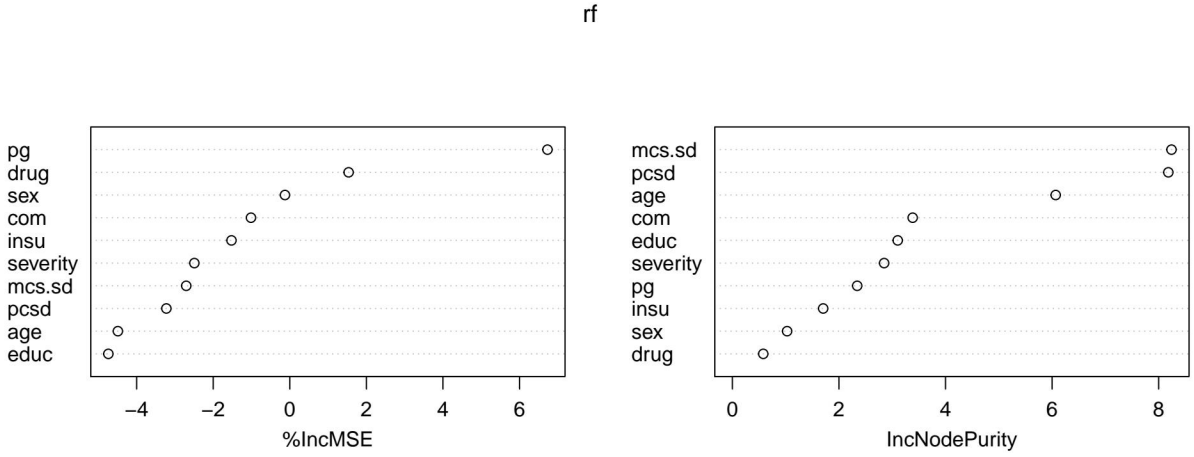
Also, with the additive model, it only explains 17.1% of deviance in training set. Moreover,from the residual plots of two randomly chosen continuous variables above,it seems the assumption of homoscedasticty is violated as the residuals deviate from the dotted line by greatly different amounts.





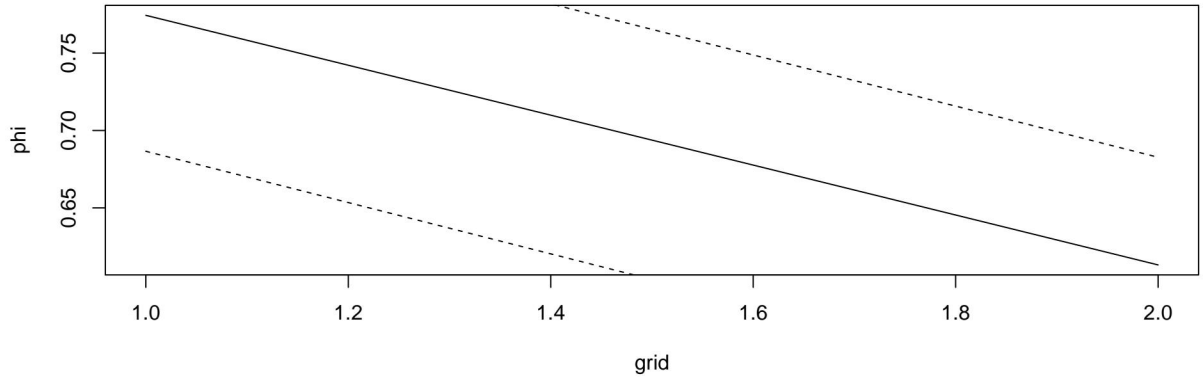
In addition, when we look at the cooks distance plots for both logistic and additive classifiers, log classifier model seems to be influenced by 3-5 data points more than the majority of data points, whereas additive model seems to be extremely influenced by 1 particular data point more than the majority of the data points.

Considering these two models' lower accuracy and sensitivity to several data points and additive model's violation of homoscedasticity assumption, I decided to stick to random forest classifier.



Now, regarding the diagnosis of our random forest, %IncMSE indicates the increase in mse of predictions as a result of certain variable being permuted(values randomly shuffled); the higher number, more important the variable is. Thus, from the visualisation above, it seems there are 3 important variables(pg,drug, sex). Secondly, IncNodePurity is the total decrease in "node impurities" from splitting on the variable, averaged over all trees. Node impurity is just residual sums of squares.It is also worth nothing that from the IncNodePurity, we know mcs.sd split reduces the node impurity the most and drug split reduces the node impurity the least.

For the second research question, I decided to use logistic regression for inferring the causal effect of physician group on patient satisfaction.



phi and grid in the visualization represent quality and pg respectively. As the line is in between the confidence interval, the causal effect of physican group on patient satisfaction is not statistically significant. Also, the overall trend shows pg 1 is better than pg 2. This causal inference is under the assumption that we have measured all confounding variables.

Now, for research question 3, we used graphical modeling to understand the relationship between the variables: pg,sex and quality.

```
##
## Call:
## glm(formula = n ~ pg * sex * quality, family = "poisson", data = df)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.03922    0.91568  -0.043 0.965835
## pg              1.64866    0.48833   3.376 0.000735 ***
## sex             2.10639    1.04167   2.022 0.043163 *
## quality         2.43159    1.03889   2.341 0.019254 *
## pg:sex         -0.82546    0.56428  -1.463 0.143505
## pg:quality     -1.09659    0.56692  -1.934 0.053076 .
## sex:quality    -0.52116    1.18504  -0.440 0.660093
## pg:sex:quality  0.42292    0.65678   0.644 0.519620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.0973e+02  on 7  degrees of freedom
## Residual deviance: 2.2204e-15  on 0  degrees of freedom
## AIC: 57.091
##
## Number of Fisher Scoring iterations: 3
```

However, as you can see from the summary above, there is no variable that seems to be statistically significant, as the p-values are all greater than 0.05. In terms of interpretation, this means there is no significant edge between pg and sex and quality. Thus, we end up with an empty graph; all the variables are independent from each other. Only if when we consider the significance level of 94%, then would we say there is a graph, in which pg and quality share a statistically significant edge.

Final Model

Before deciding on a final model, I first needed to take the strengths and weaknesses of each machine learning model into account. In general, logistic model's strengths are that they can be regularized to avoid overfitting and also be updated easily with new data using stochastic gradient descent. However, they tend to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships. On the other hand, additive model is less general than the linear model but it is much more flexible than the linear model. It is often a good compromise between the linear model and a completely nonparametric model. However, model will lose predictability when the smoothed variables' values are outside of the range of training dataset. Lastly, with random forest, they are generally good at handling unbalanced and missing data. However, their weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

In conclusion, for research question 1, we have built logistic, generalised additive and random forest models.While considering strengths and weaknesses of each model, I chose random forest to be my final choice of the model for quality prediction task because it had the lowest prediction error. Also, it's worth noting that two models are very sensitive to several data points; in addition, additive model does not meet homoscedasticity assumption well.

Then, for our 2nd research question, we used the logistic regression model with bootstrap to confirm that physician group has no causal effect on patient satisfaction under the assumption of having measured all confounding variables, as the line in the graph was within the 95% CI. Then, for our 3rd research question, we used graphical modelling to understand that there is no significant edge between pg and sex, and we subsequently end up with empty graph; all the variables are independent from each other.Only if when we consider the significance level of 94%, then would we say there is a graph, in which pg and quality share a statistically significant edge.

However, before using this model,there are several limitations that we should be aware of. This model is created using the data, which has 276 rows. In the grand scheme, the dataset is not necessarily that large; I would be cautious when applying a model trained only on 207 rows(of training set) to real application. Secondly, random forests may over-fit data sets that are particularly noisy, so it is essential that we double-check that the dataset we used to train our model is reliable.Therefore, I would recommend still being cautious when applying the following model in real-life scenarios.

3 Modeling 30 / 30

✓ **- 0 pts Correct**

- 0 pts When fitting parametric models, it would be better to mention a little bit why you didn't use higher-order terms, interaction terms, transformation, etc. Say, from EDA we don't see any particular nonlinear relationship, or I tried a few transformations but they didn't help. (For nonparametric models, no need to try transformation, etc because they would essentially transform the variables.)

- 2 pts Didn't calculate/show the confidence interval of causal effect.

- 3 pts When doing causal inference, it's important to include all possible confounding variables as the covariates, even if those variables may not be significant/do not have a clear relationship with the outcome in EDA. We may estimate the causal effect more precisely, in this way.

- 1 pts It would be better to use a test set to evaluate models that risk overfitting the data. In this dataset, training error can be very different from test error.

- 2 pts A linear model is not appropriate for binary outcome.

- 2 pts Didn't fit the counts when using the log-linear model.

- 1 pts "Most of the variables are binary" seems incorrect to me. Only three predictors, sex, pg, and drug, are binary.

- 1 pts Semiparametric models mostly refer to models that have both parametric parts and non-parametric parts, like additive models, etc. Random forests should be non-parametric models.

- 2 pts Accuracy shouldn't be below 0.5

- 1 pts I think you meant 'error' when you said "Accuracy".

- 2 pts From the first table on page 6, errors of logistic model, tree and random forest are 0.30,0.35, 0.28. Why did you say logistic model has the smallest error?

- 1 pts Prediction error shouldn't be larger than 50% if you meant the fraction of misclassified data points.

- 1 pts If the outcome is binary, linear models normally wouldn't be better than logistic models. If you are pretty sure your results are correct, please ask for a regrade.

- 3 pts Error with the plugging-in estimator. The two estimates shouldn't be the same.

- 2 pts Log-linear model is used to find the causal relationships, not to predict the outcome quality. Comparing log-linear models with logistic and additive models doesn't make sense to me.

- 5 pts Didn't use quantitative analysis (like plug-in estimator) to estimate the causal effect. Correlations are not causality

- 15 pts Sorry I didn't find your mode for goal 2 and goal 3

However, as you can see from the summary above, there is no variable that seems to be statistically significant, as the p-values are all greater than 0.05. In terms of interpretation, this means there is no significant edge between pg and sex and quality. Thus, we end up with an empty graph; all the variables are independent from each other. Only if when we consider the significance level of 94%, then would we say there is a graph, in which pg and quality share a statistically significant edge.

Final Model

Before deciding on a final model, I first needed to take the strengths and weaknesses of each machine learning model into account. In general, logistic model's strengths are that they can be regularized to avoid overfitting and also be updated easily with new data using stochastic gradient descent. However, they tend to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships. On the other hand, additive model is less general than the linear model but it is much more flexible than the linear model. It is often a good compromise between the linear model and a completely nonparametric model. However, model will lose predictability when the smoothed variables' values are outside of the range of training dataset. Lastly, with random forest, they are generally good at handling unbalanced and missing data. However, their weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

In conclusion, for research question 1, we have built logistic, generalised additive and random forest models.While considering strengths and weaknesses of each model, I chose random forest to be my final choice of the model for quality prediction task because it had the lowest prediction error. Also, it's worth noting that two models are very sensitive to several data points; in addition, additive model does not meet homoscedasticity assumption well.

Then, for our 2nd research question, we used the logistic regression model with bootstrap to confirm that physician group has no causal effect on patient satisfaction under the assumption of having measured all confounding variables, as the line in the graph was within the 95% CI. Then, for our 3rd research question, we used graphical modelling to understand that there is no significant edge between pg and sex, and we subsequently end up with empty graph; all the variables are independent from each other.Only if when we consider the significance level of 94%, then would we say there is a graph, in which pg and quality share a statistically significant edge.

However, before using this model,there are several limitations that we should be aware of. This model is created using the data, which has 276 rows. In the grand scheme, the dataset is not necessarily that large; I would be cautious when applying a model trained only on 207 rows(of training set) to real application. Secondly, random forests may over-fit data sets that are particularly noisy, so it is essential that we double-check that the dataset we used to train our model is reliable.Therefore, I would recommend still being cautious when applying the following model in real-life scenarios.

#Introduction

```
library(ggplot2)
library(mgcv)
library(tidyverse)
data<-read.table("/Users/joonghochoi/Desktop/asthma.txt",header=TRUE)
head(data,3)
```

#EDA

```
#sum(is.na(data))
cols<-character(nrow(data))
cols[]<-"black"
cols[data$sex==0]<-"red"
cols[data$sex==1]<-"blue"
ggplot(data,aes(x=quality,fill=cols))+geom_bar()+ggtitle("Histogram of quality accounting for
sex")+scale_fill_discrete(labels=c("sex=0","sex=1"))
cols[data$pg==1]<-"red"
cols[data$pg==2]<-"blue"
ggplot(data,aes(x=quality,fill=cols))+geom_bar()+ggtitle("Histogram of quality accounting for
pg")+scale_fill_discrete(labels=c("pg=1","pg=2"))

cols<-character(nrow(data))
cols[]<-"black"
cols[data$quality==0]<-"red"
cols[data$quality==1]<-"blue"
tmp<-data[,c("age","pcsd","mcs.sd")]
pairs(tmp,col=cols,pch=19,main="Pairwise plot of quantitative variables in data for EDA")
```

#Modeling
#data transformation

```
data$educ=as.factor(data$educ)
data$insu=as.factor(data$insu)
data$drug=as.factor(data$drug)
data$severity=as.factor(data$severity)
#data$com=as.factor(data$com)
data1<-data
```

#getting testing errors of different models

```
set.seed(42)
```

```
dt = sort(sample(nrow(data), nrow(data)*.75))
train<-data[dt,]
test<-data[-dt,]
library(randomForest)
rf=randomForest(quality~.,data=train,importance=TRUE)

drop <- c("quality")
testX = test[,!(names(test) %in% drop)]
testY<-test$quality
p=predict(rf,type="response",testX)
yhat=rep(0,length(testY))
yhat[p>=0.5]=1
T=table(testY,yhat)
error=(T[1,2] + T[2,1])/sum(T)
#1-error #0.6666667

log_reg= glm(quality~.,family="binomial",data=train)
p=predict(log_reg,type="response",testX)
yhat=rep(0,length(testY))
yhat[p>=0.5]=1
T=table(testY,yhat)
error=(T[1,2] + T[2,1])/sum(T)
#1-error #0.6231

add_reg<-
gam(quality~pg+s(age)+sex+educ+insu+drug+severity+com+s(pcsd)+s(mcs.sd),data=train)
p=predict(add_reg,type="response",testX)
yhat=rep(0,length(testY))
yhat[p>=0.5]=1
T=table(testY,yhat)
error=(T[1,2] + T[2,1])/sum(T)
#1-error #0.6521739



#model diagnosis
r=residuals(add_reg)
plot(train$pcsd,r)
plot(train$mcs.sd,r)
summary(add_reg) #17.1%

plot(cooks.distance(log_reg),type="h")
plot(cooks.distance(add_reg),type="h")
varImpPlot(rf)
```

```
#Causal inference

#rf1=randomForest(quality~.,data=data1,importance=TRUE)
log_reg= glm(quality~.,data=data)
n=2
grid=seq(1,2)
phi=rep(0,n)
for (i in 1:n){
 pg1=grid[i]
 new_df=data1
 new_df$pg=pg1
 phi[i]=mean(predict(log_reg,new_df,type="response"))
 #phi[i]=mean(predict(rf1,new_df,type="response"))
}

N=2
#N_bootstrap=100
N_bootstrap=1000
grid=seq(1,2)
n=length(data$quality)
phi_temp=matrix(0,N_bootstrap,N)

set.seed(42)
for (i in 1:N_bootstrap){
 I=sample(1:n,size=n,replace=TRUE)
 bootstrap_df=data1[I,]
 #bootstrap_out=randomForest(quality~.,data=bootstrap_df,importance=TRUE)
 bootstrap_out=glm(quality~.,family="binomial",data=bootstrap_df) #try different
classifiers
 bootstrap_new_df=bootstrap_df
 for (j in 1:N){
  set_pg=grid[j]
  bootstrap_new_df$pg=set_pg
  phi_temp[i,j]=mean(predict(bootstrap_out,bootstrap_new_df,type="response"))
 }
}

l=rep(0,N)
u=rep(0,N)
for (i in 1:N){
 phi_temp_sort=sort(phi_temp[,i])
 #l[i]=phi_temp_sort[2]
 #u[i]=phi_temp_sort[97]
 l[i]=phi_temp_sort[25] #2.5% of the bootstrap
 u[i]=phi_temp_sort[976] #97.75% of the bootstrap in integer
}
```

```
plot(grid,phi,type='l')
lines(grid,l,type='l',lty=2)
lines(grid,u,type='l',lty=2)

#Graphical modeling

a<-data%>%filter(sex==0,quality==0)%>%group_by(pg)%>%summarize(count=n())
b<-data%>%filter(sex==1,quality==0)%>%group_by(pg)%>%summarize(count=n())
c<-data%>%filter(sex==0,quality==1)%>%group_by(pg)%>%summarize(count=n())
d<-data%>%filter(sex==1,quality==1)%>%group_by(pg)%>%summarize(count=n())
n<-c(a$count,b$count,c$count,d$count)
pg<-c(1,2,1,2,1,2,1,2)
sex<-c(0,0,1,1,0,0,1,1)
quality<-c(0,0,0,0,1,1,1,1)
df=data.frame(pg=pg,sex=sex,n=n,quality=quality)

out = glm(n~pg*sex*quality,data=df,family="poisson")
summary(out)
```

4 Final Models 25 / 25

✓ - **0 pts** **Correct**

- **2 pts** Conclusions should contain a more elaborate discussion of the results of the final models, trying to connect them to the actual data and the problem you were required to solve. For instance, for the causal inference part you could have told more about why you think your model did or did not uncover any causal effect, commenting on possible omitted-variable bias or on the plausibility in a real-world setting of the effect you found, if any.
- **5 pts** No conclusion about predictive model
- **5 pts** No conclusions about causal inference
- **5 pts** No conclusions about graphical models
- **8 pts** No limitations
- **3 pts** Taking the average of training and test error is wrong. Training error underestimates the true generalization error, hence if you had a model that strongly overfits the data you might end up with a suboptimal choice
- **4 pts** Insufficient explanation in limitations
- **8 pts** No discussion of results, just reported some metrics
- **25 pts** Missing entirely