Unexpected Contributors of COPD

Brian Dong

STAT610

Fall 2021

12/8/2021

Abstract:

It’s well known that for COPD (chronic obstructive pulmonary disease) it manifests primarily from symptoms such as emphysema. Most people generally understand that damaged lung tissue comes from exposure to smoke and other chemical pollutants which influences the bulk of problems in the lungs. What people don’t realize is the extent other factors pre existing conditions or current therapy and medications affect the extent of damaged lung tissue. Overall we come to understand that while some areas such as gender, supplemented oxygen, gender and race or even bmi have significant roles on the extent of emphysema affected tissue, other areas such as height do not.

Intro:

COPD relates the trapping of airflow in the lungs. It is well known that COPD is caused by people generally diagnosed with emphysema that involve damage to the alveoli tissue leading to an inability to support bronchial tubes causing air to be trapped in the lungs. While it is generally known that smoking and other pollutants coming into contact with the lungs causes this damage, it is still not well known how significant other factors may influence the extent of emphysema. Are there other variables that influence the damage of lung tissue outside of smoking? Variables such as pneumonia, race, gender, O2 hours per day, age, and height are all worth examining in patients with COPD especially since they can be related to pre existing problems before they acquired or were diagnosed with COPD. So our hypothesis is are these non-smoking related variables significant as predictors of lung tissue damage? This project will use a multivariate linear regression to analyze the response which is a quantitative value predicted off both category and numerical variables or quality and quantity variables.

Exploratory Data Analysis

1) Data Tables and variable choice

The data set we will use includes 35 variables and 5747 patients. We can see a head preview of our dataset from the top. We would like to examine continuous variables pct_gastrapping, weight_kg, height_cm, BMI, and age at time of visit. We will also examine categorical variables as well including pneumonia (yes or no), race, and gender.

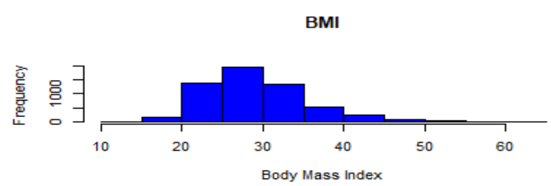
```
     sid visit_year visit_date visit_age gender  race height_cm weight_kg sysBP diasBP hr O2_hours_day   bmi  asthma
1 10005Q       2008  1/15/2008      54.5 Female White     159.9      73.0   130     80 87            0 28.55      No
2 10006S       2008  1/15/2008      62.3 Female White     162.6      86.0   170     80 81            8 32.53      No
3 10010J       2008  1/15/2008      65.9 Female White     162.1      62.8    96     63 66            0 23.90      No
4 10015T       2008  2/15/2008      59.6   Male White     182.9     110.0   142     88 75            0 32.88     Yes
5 10017X       2008  6/15/2008      67.5   Male White     179.1      83.0   106     72 72           10 25.88 unknown
6 10022Q       2008  2/15/2008      69.8 Female White     158.8      78.0   122     78 87            0 30.93      No
  hay_fever bronchitis_attack pneumonia chronic_bronchitis emphysema copd sleep_apnea SmokStartAge CigPerDaySmokAvg
1         0                No        No                 No        No   No          No           14               20
2         0               Yes       Yes                 No       Yes  Yes          No            8               20
3         0           unknown       Yes                Yes        No  Yes          No           25               15
4         1           unknown       Yes            unknown   unknown  Yes         Yes           16               20
5         0               Yes       Yes                 No       Yes  Yes          No           20               40
6         3           unknown       Yes            unknown       Yes  Yes          No           13               30
  Duration_Smoking smoking_status total_lung_capacity pct_emphysema functional_residual_capacity pct_gastrapping
1             40.5 Current smoker              5.6636      0.926851                       2.4766         6.80077
2             52.0  Former smoker              5.2325     14.005900                      -1.0000        -1.00000
3             40.9 Current smoker              5.1960      1.683760                       3.8993        41.34930
4             28.0  Former smoker              6.3971      9.330450                      -1.0000        -1.00000
5             35.0  Former smoker              7.8935     36.262400                       4.1043        46.17690
6             30.0  Former smoker              5.1016     30.484400                      -1.0000        -1.00000
  insp_meanatt exp_meanatt FEV1_FVC_ratio  FEV1   FVC FEV1_phase2
1     -830.343    -650.526           0.77 2.921 3.805       2.622
2     -841.880      -1.000           0.43 1.288 3.022       1.318
3     -833.429    -789.595           0.53 1.008 1.909       1.087
4     -841.315      -1.000           0.51 1.906 3.732       2.002
5     -887.947    -792.397           0.57 2.748 4.827       2.178
6     -865.608      -1.000           0.53 1.076 2.047       0.924
> 
```

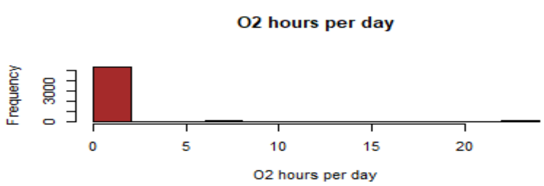
2) Visualizing Predictors and response

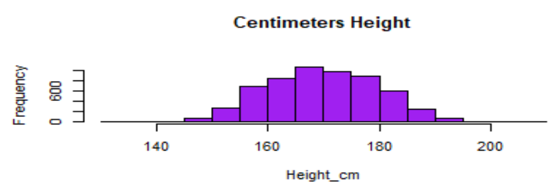
i) Histograms and variable conversions

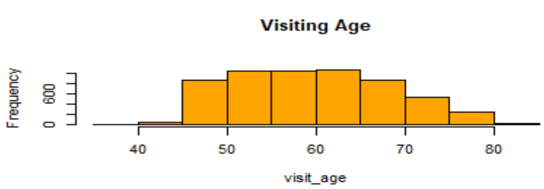
We have to convert categorical variables into a numerical form since certain plots later on can only show numerical values. In the case of the categorical values becoming numeric we present it as 1 = No, 2 = Unknown, and 3 = Yes.

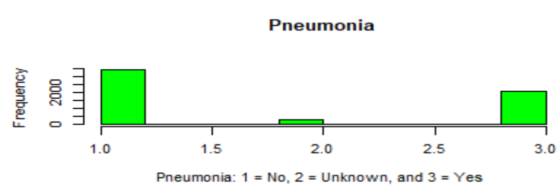
Histograms provide a useful way to summarize our discrete and continuous data. Our histograms can show the degree of skewness for our current data.Looking at the first four histograms for our continuous variables, we can see that our response variable for percent emphysema is heavily right skewed along with O2_hours_day, and BMI is partially skewed to the right, while height in centimeters and visiting age are reasonably normally distributed. We also have histograms that use our converted categorical variables. We have an almost equal balance of female and male patients while there are a large number of patients with pneumonia with a small amount having unknown status. There are also clearly much less black patients than white patients in this data collection.

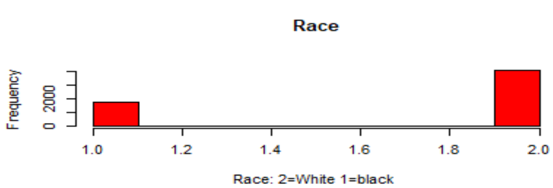


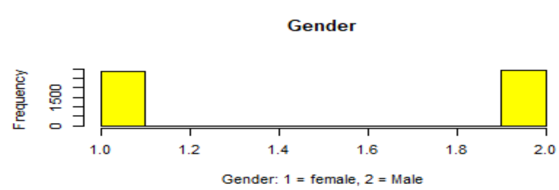












3) Pairing variables and correlations

We don't necessarily have to scale our values since the range of values we are given for each variable is the same. We end up plotting our predictor variables and our response variable on a ggpairs plot. The plot is useful as it provides a way for us to see correlation among paired variables. We can see that there is so far no signs of multicollinearity among any of the variables. We see a moderate level of correlation between visiting age and percentage of emphysema. Just as we saw in our histogram above there is a big difference in regards to race. Among these plots we do not see any serious signs of multicollinearity.

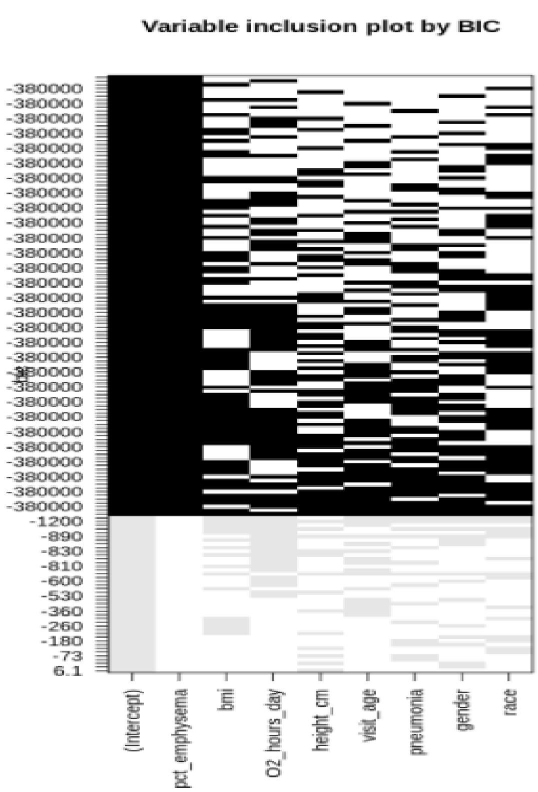
Statistical Analysis

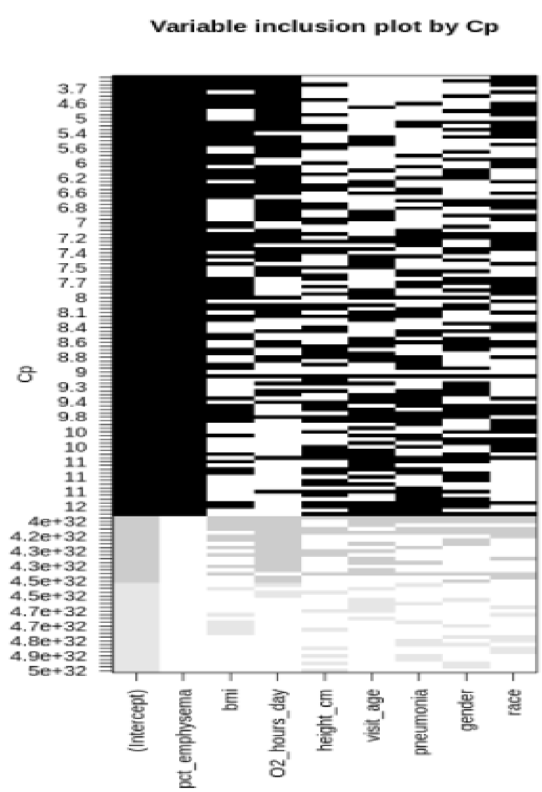
1) Addressing transformations and other data

Earlier we ended up seeing that our data for the response variable percent emphysema had a strong right or positive skew. In our model later it is important to remember that when looking at the error residuals we are looking for the level of variance to be homogenous. Therefore we will choose to .25 square root the response values. Beyond this we do not have to transform anymore of the data unless the other assumptions for errors in our residual model. Furthermore we do not have missing data that has to be removed or deleted from our variables.

2) Variable Selection

Before we get a final model we will have to perform subset analysis. Fortunately we don't have the largest amount of parameters so it is fine to choose between BIC and Cp (a variant of AIC). While both assess model fit based on penalized model parameters and checking overfitted data AIC tends to prefer a more complex model compared to BIC which selects a more simple model which we will look at. Regardless of the models, the inclusion plots are comparing the models based on each row which includes the negative sections and the black lines indicate variable inclusion within the model. So based on both, they include **O2_hours_day, visit_age, gender, and bmi**.





3) Final model and Diagnostics

We run our final model after selecting from BIC and look at the parameters on our model. The overall model was significant with a p value of almost 0, and we have our adjusted R^2 value which is 0.2491. All our predictors were significant around the α=0.5 level except height_cm. It seems that O2_hours_day, visit_age, gender, and bmi were all relatively around the same in contributing toward significance of this model. Pneumonia and race are somewhat less significant

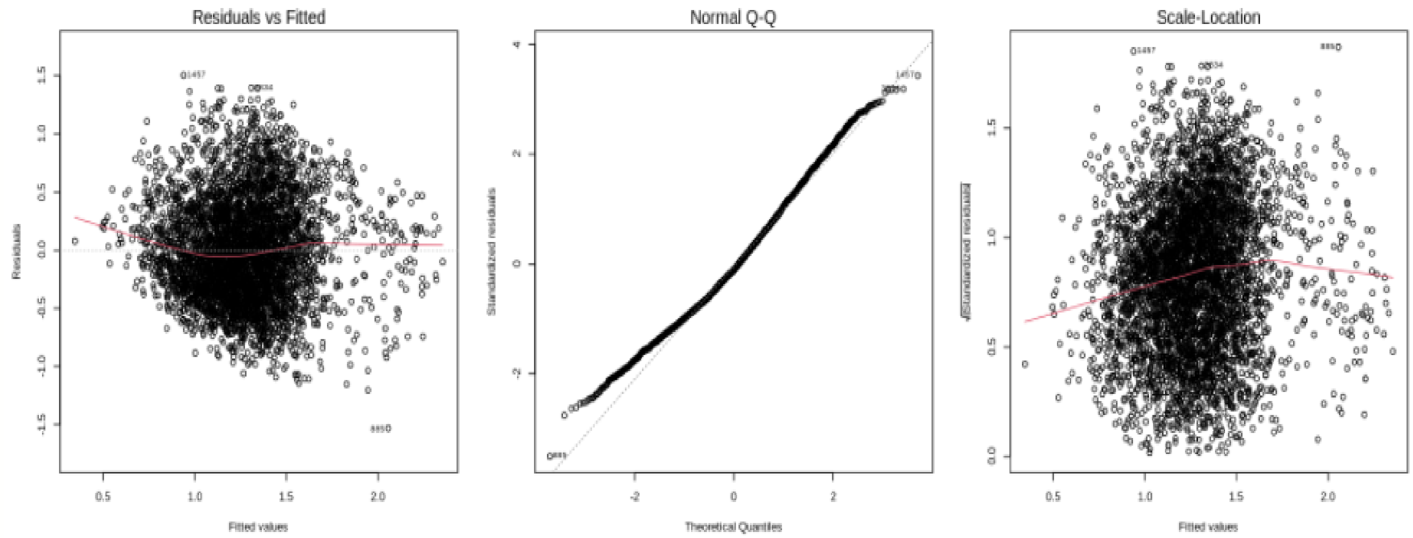
```
Call:
lm(formula = pct_emphysema^0.25 ~ bmi + O2_hours_day + height_cm +
    visit_age + pneumonia + gender + race, data = copd)

Residuals:
     Min       1Q   Median       3Q      Max
-1.53215 -0.32248 -0.04502  0.28620  1.50533

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1485078  0.1608798   0.923    0.356
bmi          -0.0152837  0.0010661 -14.336  < 2e-16 ***
O2_hours_day  0.0284377  0.0015471  18.381  < 2e-16 ***
height_cm     0.0009990  0.0009539   1.047    0.295
visit_age     0.0134188  0.0008173  16.419  < 2e-16 ***
pneumonia     0.0432174  0.0068691   6.292 3.43e-10 ***
gender        0.1988382  0.0181761  10.940  < 2e-16 ***
race          0.1057098  0.0156730   6.745 1.72e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4389 on 4694 degrees of freedom
  (1045 observations deleted due to missingness)
Multiple R-squared:  0.2502,    Adjusted R-squared:  0.2491
F-statistic: 223.8 on 7 and 4694 DF,  p-value: < 2.2e-16
```

After running our diagnostic plots, we can see that there are no serious violations concerning error regression assumptions. The trend line for the residuals vs fitted is fairly linear and the data falls well clustered around it. The QQ plot has most of its data on the line although the left tail is slightly left skewed. Meanwhile the Scale location plot has a relatively horzantel x=y line with a slit bump in the middle, but most of the data is heavily condensed around it showing there is no violation of homoscedasticity.



4) Inference and Model Description

Let X_1 be bmi, X_2 be O2_hours_day, X_3 be height_cm, X_4 be visit_age, X_5 be pneumonia, X_6 be gender, and X_7 be race. Our final model is this:

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 \ldots B_7X_7 + \epsilon$$

The test hypotheses we use is:

H_0 **: All the** B_i **are equal to zero**

H_α **: At least one of the** B_i **are equal to zero**

We have already examined our model and know that it is significant at the .05 significance level. Therefore we can reject our null hypothesis since at least one of the coefficients is not equal to 0. Now we have our regression line as:

$$\hat{Y} = 0.14850 + (-0.01528)X_1 + 0.02843X_2 + 0.00099X_3 + 0.01341X_4 + 0.04321X_5 + 0.19883X_6 + 0.1057X_7$$

We should also run follow up tests with confidence intervals to validate our results. Looking at the variables only height_cm includes 0 in its interval with respect to the 95% confidence interval. The other variables do not include zero. This means that this height_cm compared to the other ones is not statistically significant. We interpret some of these intervals in the following way: for example for O2_hours_day we are 95% confident that for every additional hour of oxygen supplemented, the percentage of damaged tissue area increases between .0254 and .0314.

For every kg per meter squared the percentage of damaged tissue area decreases between -0.0173 and -0.0131. For every additional year of a visitor's age the percentage of damaged tissue increases between .01181 and 0.01502. For every patient with pneumonia there is an increase between .0297 and .0566, and this follows similarly with gender and race.

```
> confint(fit)
                    2.5 %       97.5 %
(Intercept)  -0.1668921690  0.463907789
bmi          -0.0173736744 -0.013193652
o2_hours_day  0.0254045525  0.031470788
height_cm    -0.0008710066  0.002869001
visit_age     0.0118166145  0.015021031
pneumonia     0.0297506135  0.056684117
gender        0.1632044771  0.234471919
race          0.0749834049  0.136436295
> 
```

Conclusion

O2_hours_day, visit_age, gender, and bmi were the most largest predictors in regards to percentage of emphysema. These predictors did not have a standout value in terms of significance. It was expected that O2 hours per day would have more standout in terms of significance compared to the other three given the highest correlation but it didn’t. Obviously having additional oxygen in your lungs can help prevent the extent of damaged tissue and people who have taken it before they developed symptoms of emphysema may have better protection than those who don’t. Pneumonia is important as it could further increase damage to your lungs as it in its natural form is a respiratory illness. Race and Gender are also significant because different demographics of people in America have different lifestyles, meaning for example white people may smoke more cigarettes on average or black people may exercise less which influences the extent of damage on lung tissue. Females for instance have different body structures which affect the way they may breath, in fact black women are among some of the more likely to develop emphysema. BMI definitely correlates with how people may have emphysema as certain skinnier patients will have thinner walls causing stronger levels of vibration meaning more tissue damage. Height_cm was seen to be non significant with respect to the rest of the variables. Height is something that affects a person's diaphragm or size of their lungs to hold air, but it doesn’t necessarily influence tissue damage or repair. In the future we may want to use more variables to add to our model as some were disappointingly expected such as height. Being aware of current or previous medication or exercise is important in assessing these individuals that currently are having these problems as they could affect how badly their lungs could be damaged when diagnosed with COPD or emphysema.

References

1) Smith, B. M. (2018, October 24). Impact of pulmonary emphysema on exercise capacity and its physiological determinants in chronic obstructive pulmonary disease. Nature. https://www.nature.com/articles/s41598-018-34014-5?error=cookies_not_supported&code=7010ad69-9733-4d95-813f-c6664f48c680

2) *Emphysema - Symptoms and causes*. (2017, April 28). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/emphysema/symptoms-causes/syc-20355555

3) Divo, M. J., MD. (2014, July 24). *Comorbidity Distribution, Clinical Expression and Survival in COPD Patients with Different Body Mass Index*. COPD Foundation. https://journal.copdfoundation.org/jcopdf/id/1036/Comorbidity-Distribution-Clinical-Expression-and-Survival-in-COPD-Patients-with-Different-Body-Mass-Index

Appendix

```
#List of packages and libraries installed
install.packages("GGally")
install.packages("car")
install.packages("multcomp")
install.packages("caret")
install.packages("leaps")
library(ggplot2)
library(dplyr)
library(GGally)
library(car)
library(multcomp)
library(leaps)
library(caret)

#Display data table preview
copd <- read.csv("C:/Users/Admin/Dropbox/My PC
(DESKTOP-3GJ696L)/Documents/SDSU_Graduate/STAT610_Linear_Regression_Models/copd_da
ta.csv", header=TRUE)
head(copd)

#Convert categorical into numeric
copd$pneumonia=factor(copd$pneumonia)
copd$pneumonia=as.numeric(copd$pneumonia)
copd$gender=factor(copd$gender)
copd$gender=as.numeric(copd$gender)
copd$race = factor(copd$race)
copd$race = as.numeric(copd$race)


#Create histograms to show distribution of data and frequency of categorical variables
par(margin(1,1,1,1))
par(mfrow = c(4,2))


hist(copd$pct_emphysema, xlab="%emphysema", main = "Percenent of Emphysema",
   col = "pink")

hist(copd$bmi, xlab = "Body Mass Index", main = "BMI", col = "blue")

hist((copd$O2_hours_day), xlab = "O2 hours per day", main = "O2 hours per day",
   col="brown")
```

```
hist(copd$height_cm, xlab = "Height_cm", main = "Centimeters Height",
    col="purple")

hist(copd$visit_age, xlab = "visit_age", main = "Visiting Age", col = "orange")

hist(copd$pneumonia, xlab = "Pneumonia: 1 = No, 2 = Unknown, and 3 = Yes ",
    main = "Pneumonia", col= "green")

hist(copd$race, xlab = "Race: 2=White 1=black ", main = "Race", col = "red")

hist(copd$gender, xlab = "Gender: 1 = female, 2 = Male", main = "Gender",
    col = "yellow" )

copd$gender=factor(copd$gender)
copd$pneumonia=factor(copd$pneumonia)
copd$race = factor(copd$race)



#Create GGpair plot to visualize correlations
options(repr.plot.width=10, repr.plot.height=10)
options(warn=-1)
X11()
plot_frame <- data.frame("pct_emphysema" = copd$pct_emphysema, "BMI" = (copd$bmi),
"O2_Hours_per_day" = copd$O2_hours_day, "Height_cm" = copd$height_cm, "Visiting_Age" =
copd$visit_age, "Pneumonia" = copd$pneumonia,  "Gender" = copd$gender,
                    "Race"=copd$race)
ggpairs(plot_frame, aes(color=Race, alpha=100), binwidth=30)

#Convert categorical variables back into numeric values
copd$pneumonia=factor(copd$pneumonia)
copd$pneumonia=as.numeric(copd$pneumonia)
copd$gender=factor(copd$gender)
copd$gender=as.numeric(copd$gender)
copd$race = factor(copd$race)
copd$race = as.numeric(copd$race)

#Setting up predictors for subset analysis
preds = with(copd, cbind(pct_emphysema, bmi, O2_hours_day, height_cm, visit_age, pneumonia,
gender, race))

model = regsubsets(preds, y = copd$pct_emphysema,
               nbest = 30,    # save the best # for each number of variables
               nvmax = 20,    # maximum number of variables allowed in the model
```

```
                    really.big=T)  # for larger datasets

#Plot BIC vs Cp model plots
par(mfrow = c(1, 2))
plot(model, scale = "bic", main = "Variable inclusion plot by BIC")
plot(model, scale = "Cp", main = "Variable inclusion plot by Cp")

#Final Regression model and summary
fit <- lm(pct_emphysema^.25 ~ bmi+ O2_hours_day+ height_cm+ visit_age+ pneumonia of gender
race, data=copd)
summary(fit)

#Plots for residual error assumptions
options(repr.plot.width=15, repr.plot.height=3)
par(mfrow=c(1,3))
plot(fit, which=c(1,2,3))

#Confidence Interval
confint(fit)
```