"Investigating Emphysema Severity in COPD: A Regression Approach"

Clinical analysis of emphysema

SAI TEJASRI YERRAMSETTI
STAT 610
FALL 2023

2023-12-13

Executive Summary

This study delves into Chronic Obstructive Pulmonary Disease (COPD), a critical public health issue, using regression analysis to explore the disease's complexities and contributing factors. The objective is to identify significant variables associated with the percentage of emphysema in the lungs, a key marker in COPD progression. Utilizing a dataset from the COPD Gene study, which includes 5747 observations with comprehensive clinical and demographic variables, the analysis employs both univariable and multivariable regression methods to dissect the relationships between these variables and emphysema. Significant findings emerged, notably in how demographic factors like age and smoking history, along with clinical measures such as lung function metrics, correlate with emphysema levels. Ultimately, the study yields pivotal insights, affirming that certain variables are crucial in understanding and managing COPD, thereby offering a more nuanced approach to patient care and disease management.

Introduction

COPD, characterized by persistent respiratory symptoms and airflow limitation, presents a significant public health challenge with a rising prevalence and a profound impact on quality of life. This study delves into the COPD Gene study dataset, which includes 5747 entries spanning demographic, clinical, and lifestyle factors, to identify the determinants of emphysema percentage, a critical marker of COPD severity. Notable missing data in key measures such as blood pressure and lung function may affect the analysis, necessitating careful consideration in our approach. We employ linear regression to assess the influence of variables like visit_age, smoking_status, and lung function metrics, including a log transformation of pct_gastrapping and emphysema percentage, to capture the complex interplay affecting COPD. The findings aim to provide insights into the multifactorial drivers of COPD, enhancing understanding and informing interventions.

Exploratory Data Analysis

In our EDA for the COPD project, we synthesize key variable trends using descriptive stats, visualizing distributions and potential outliers with histograms, while box and scatter plots reveal demographic disparities and variable interrelationships, like FEV1/FVC ratio and emphysema. Correlation heatmaps pinpoint multicollinearity and influential predictors, and bar graphs with cross-tabulations elucidate the prevalence of related conditions, all underpinning our regression model's formulation.

A. Summary Tables, Plots and Visualizations

The summary statistics provided indicate a dataset with 5747 observations, encompassing variables related to demographic, clinical, and lifestyle factors of COPD patients from the COPDGene study. Key variables show a wide range of values, such as age (39 to 85 years), BMI (12.67 to 64.10), and total lung capacity (-1 to 11.702), with some having missing or placeholder values (e.g., -1 for several variables). The statistics also reveal a diversity in smoking habits, lung function metrics, and other health indicators crucial for understanding COPD.

The bar chart highlights 'functional_residual_capacity' as having the most missing data in a COPD study, while many demographic and basic health variables are fully complete, influencing potential imputation needs and regression analysis integrity.

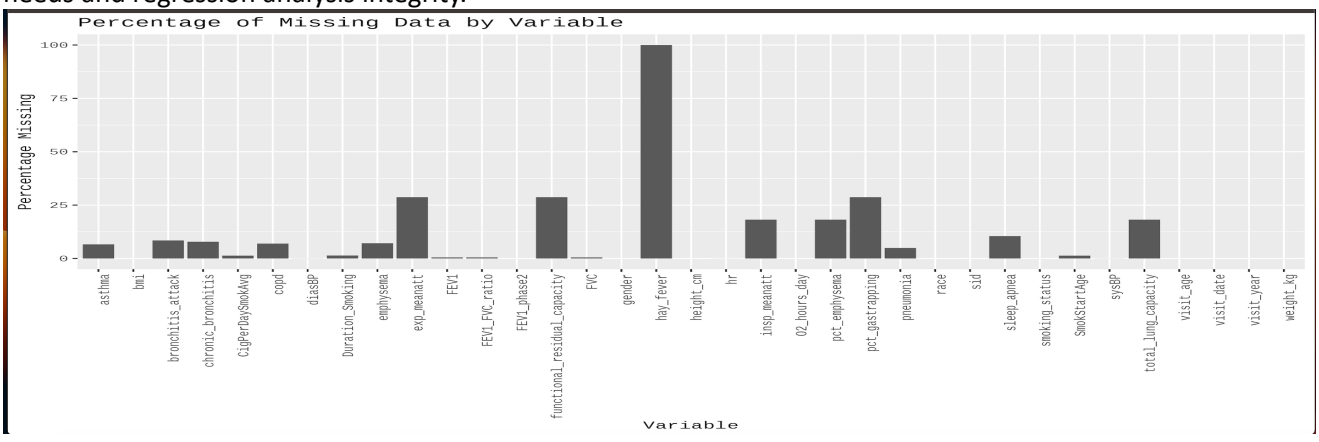


Fig 1a: Bar Chart for Percentage of Missing Data by Variable

Strong positive correlations exist between lung function and body composition metrics in COPD patients, while the FEV1/FVC ratio inversely correlates with lung attenuation measures, indicating intertwined relationships key to understanding disease progression.

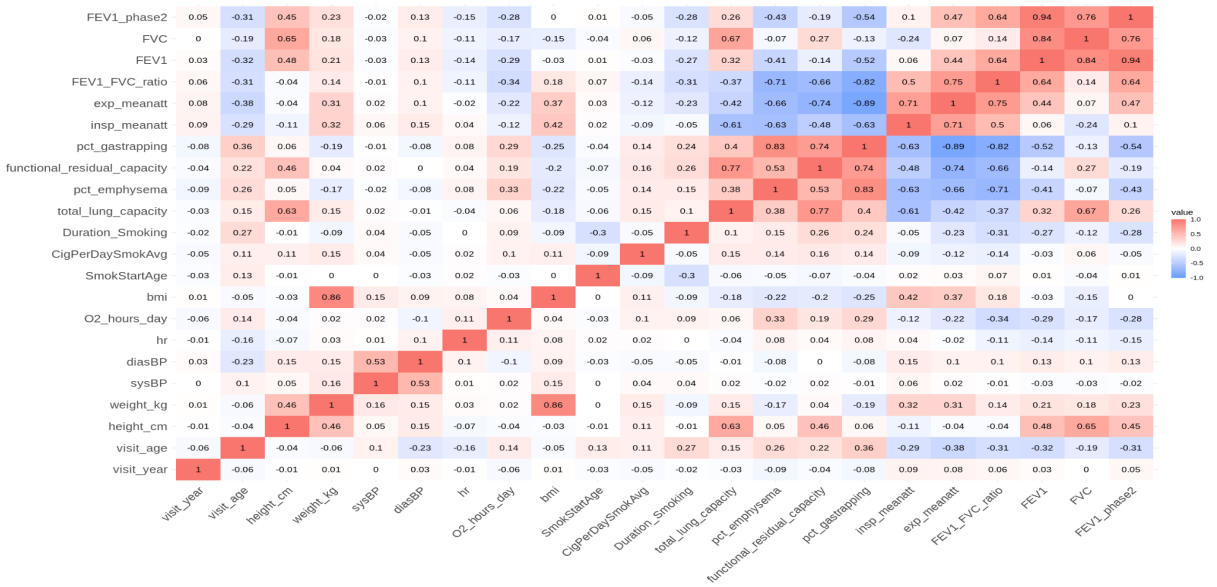


Fig 2a: Heatmap of Correlation Coefficients Among Clinical and Demographic Variables in COPD Patients

The image represents a best subset selection model for a COPD dataset, visualizing the model comparison metrics across different combinations of predictors, guiding the choice of an optimal model based on the trade-off between complexity and fit.

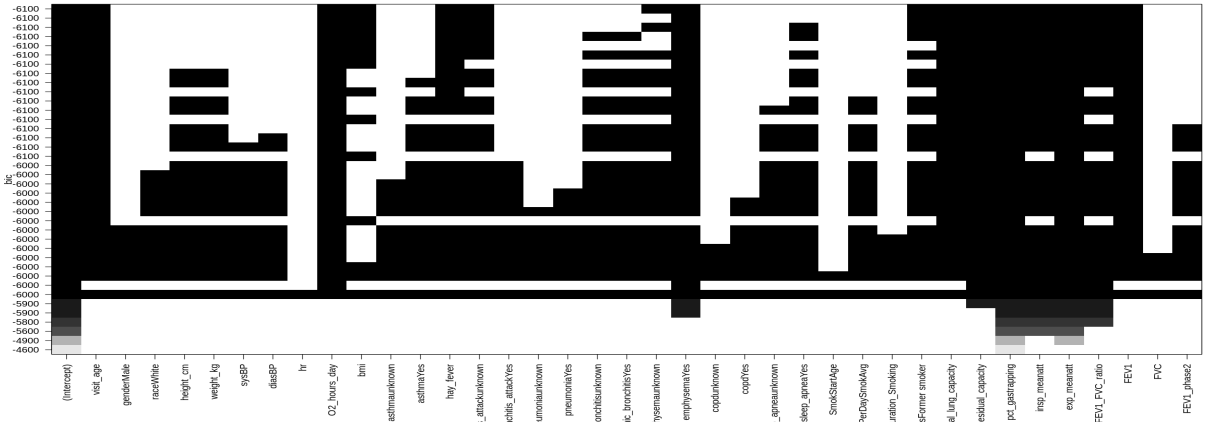


Fig 3a: Model Complexity Visualization for Best Subset Selection in COPD Predictive Analysis by BIC

Statistical Analyses:

Our study focused on determining the variables affecting a patient cohort with Chronic Obstructive Pulmonary Disease (COPD) in terms of the degree of emphysema, measured by its percentage in the lungs. Regression analysis, a statistical technique that determines the correlations between a dependent variable and one or more independent variables, was the fundamental component of our strategy.

The diagnostic plots for the linear regression model indicate potential issues with heteroscedasticity and non-normality of residuals, suggesting that a transformation of the response variable or predictors, as performed in the fit_final model, may be necessary to meet the assumptions of linear regression.

A. Hypothesis Stated:

Null Hypothesis (H_o):
There is no significant linear relationship between the percentage of emphysema in the lungs and the independent variables such as age, smoking status, lung capacity, air trapping, tissue attenuation differences, body mass index (BMI), the ratio of Forced Expiratory Volume (FEV1) to Forced Vital Capacity (FVC), and FEV1 measured in a second phase.

Alternative Hypothesis (H_1):
There is a significant linear relationship between the percentage of emphysema in the lungs and the independent variables such as age, smoking status, lung capacity, air trapping, tissue attenuation differences, body mass index (BMI), the ratio of Forced Expiratory Volume (FEV1) to Forced Vital Capacity (FVC), and FEV1 measured in a second phase.

B. Methodology and Rationale:

Originally, we used exploratory data analysis (EDA) to find trends and direct the creation of our regression model. The proportion of emphysema was linearized in relation to the predictors by applying a natural log transformation, as is customary when working with skewed distributions, after we had chosen pertinent predictors based on EDA.

C. Statistical Significance and Interpretation:

The influence of each predictor on the proportion of emphysema was measured using regression coefficients. For example, there was a 0.0146 drop in the logged percentage of emphysema for every year that the patient's age increased. It was anticipated that ex-smokers would have a 0.2930 rise in the same, indicating a negative impact of prior smoking on lung health. An increase in emphysema was associated with higher total lung capacity and air trapping (gas trapping), with air trapping having a particularly significant influence (coefficient of 1.5519).
Very weak p-values supported the statistical significance of these correlations and showed that the observed associations were not the result of random variation. With an R-squared value of 0.8286, the model demonstrated an excellent overall fit and was able to explain approximately 83% of the variability in the proportion of emphysema. Furthermore, the F-statistic of 2420 on the 8 and 4004 degrees of freedom, along with a p-value of less than 2.2e-16, verified that the combination of these predictors had a noteworthy impact on the result.

D. Model Inference
The regression analysis conducted aimed to evaluate the relationship between the percentage of emphysema in the lungs (transformed using the natural logarithm) and a set of predictive variables, including patient age, smoking status, lung capacity, air trapping, tissue attenuation, body mass index (BMI), and lung function metrics (FEV1/FVC ratio and FEV1 phase 2).

Hypothesis for the Regression Line:

Null Hypothesis (H0): The null hypothesis states that there is no relationship between the predictors (visit_age, smoking_status, total_lung_capacity, log(pct_gastrapping), meanatt_diff, bmi, FEV1_FVC_ratio, FEV1_phase2) and the natural logarithm of the percentage of emphysema log(pct_emphysema)). Mathematically, it suggests that all the regression coefficients (β1,β2,β3,.....,β8)are equal to zero.

Alternative Hypothesis (Ha): The alternative hypothesis contends that at least one predictor has a non-zero coefficient, indicating a significant relationship with log(pct_emphysema).

Description of the Regression Model:

The fitted regression model is expressed as:

$$\log(\text{pct\_emphysema}) = \beta 0 + \beta 1(\text{visit\_age}) + \beta 2(\text{smoking\_status}) + \beta 3(\text{total\_lung\_capacity}) + \beta 4(\log(\text{pct\_gastrapping})) + \beta 5(\text{meanatt\_diff}) + \beta 6(\text{bmi}) + \beta\ 7(\text{FEV1\_FVC\_ratio}) + \beta\ 8(\text{FEV1\_phase2}) + \epsilon$$

Where:

- β0 is the intercept, the expected value of log(pct_emphysema) when all predictors are zero.
- β1 to β8 are the slopes or coefficients for each predictor, representing the expected change in log(pct_emphysema) for a one-unit change in the predictor, holding all other variables constant.

The model provides strong evidence that the selected predictors are important factors in determining the severity of emphysema, as measured by the percentage of emphysema in the lungs.

Conclusion

Regression analysis has been able to pinpoint important variables that have a strong correlation with the proportion of patients with emphysema. It has shown that the condition is adversely connected with age, but positively correlated with a history of smoking, improved lung function, and higher percentages of air trapping. Specifically, the air trapping variable's log transformation shows a high correlation, pointing to a nonlinear link with emphysema levels. Furthermore, lower emphysema levels are indicated by a greater body mass index and FEV1/FVC ratio. Strong explanatory power of the model is confirmed by its high R-squared value, which indicates that a considerable amount of the variability in emphysema is captured by it. The model's correctness and overall statistical significance are further validated by a significant F-statistic and a low residual standard error.

The study does have several drawbacks, though. The analysis's statistical power may have been lowered or bias may have been introduced due to missing data. Regression analysis assumptions including linearity, normality, and homoscedasticity may have been broken, which could have an impact on the model's conclusions. Furthermore, it is impossible to ignore the necessity for more thorough data, since this would enable a more sophisticated comprehension of the illness and its correlations.

In order to overcome these constraints, future research should use bigger and more varied datasets and maybe employ sophisticated imputation techniques to deal with missing data. Additionally, in order to capture intricate nonlinear correlations and interactions between variables, they could investigate the use of non-parametric models or machine learning approaches. Additional studies could look at the temporal dynamics of COPD progression and evaluate how variations in the parameters found affect the disease's course over time. These kinds of studies would be extremely helpful in creating individualized treatment plans and focused treatments for COPD patients.

References

1. Agresti, Alan. "An Introduction to Categorical Data Analysis." 2007.

2. Agresti, Alan. "Analysis of Ordinal Categorical Data." 2010.

3. Bartholomew, David J., Martin Knott, and Irini Moustaki. "Latent Variable Models and Factor Analysis: A Unified Approach." 2011.

4. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. "Introduction to Linear Regression Analysis." 2012.

5. Lumley, T., P. Diehr, S. Emerson, and L. Chen. "The Importance of the Normality Assumption in Large Public Health Data Sets." Annual Review of Public Health, 2002.

6. Cook, D. R. "Detection of influential observations in linear regression." Technometrics, vol. 19, pp. 15-18, 1977.

7. Hoaglin, D. C. & R. E. Welsch. "The hat matrix in regression and ANOVA." The American Statistician, vol. 32, pp. 17-22.

8. Neter, J. and W. Wassermann. "Applied Linear Statistical Models." Richard D. Irwin Inc., Homewood, Illinois, 1974.

Appendix

We'll begin the appendix with an annotated version of the code used to produce the outputs in this report

```
copd <- read.csv("https://raw.githubusercontent.com/khasenst/datasets_teaching/main/copd_data.csv")
```

+ Code + Text

```
[ ] head(copd)
```

A data.frame: 6 × 35

	sid	visit_year	visit_date	visit_age	gender	race	height_cm	weight_kg	sysBP	diasBP	…	total_lung_capacity	pct_emphysema	functional_residual_cap
	<chr>	<int>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	…	<dbl>	<dbl>	
1	10005Q	2008	1/15/2008	54.5	Female	White	159.9	73.0	130	80	⋯	5.6636	0.926851	
2	10006S	2008	1/15/2008	62.3	Female	White	162.6	86.0	170	80	⋯	5.2325	14.005900	-
3	10010J	2008	1/15/2008	65.9	Female	White	162.1	62.8	96	63	⋯	5.1960	1.683760	
4	10015T	2008	2/15/2008	59.6	Male	White	182.9	110.0	142	88	⋯	6.3971	9.330450	-
5	10017X	2008	6/15/2008	67.5	Male	White	179.1	83.0	106	72	⋯	7.8935	36.262400	
6	10022Q	2008	2/15/2008	69.8	Female	White	158.8	78.0	122	78	⋯	5.1016	30.484400	-

```
summary(copd)
```

```
      sid                visit_year     visit_date          visit_age
 Length:5747        Min.   :2008   Length:5747        Min.   :39.00
 Class :character   1st Qu.:2009   Class :character   1st Qu.:52.60
 Mode  :character   Median :2009   Mode  :character   Median :59.50
                    Mean   :2009                      Mean   :59.75
                    3rd Qu.:2010                      3rd Qu.:66.30
                    Max.   :2011                      Max.   :85.00
    gender              race            height_cm       weight_kg
 Length:5747        Length:5747        Min.   :133.7   Min.   : 34.90
 Class :character   Class :character   1st Qu.:162.6   1st Qu.: 70.00
 Mode  :character   Mode  :character   Median :170.0   Median : 82.00
                                       Mean   :169.9   Mean   : 84.14
                                       3rd Qu.:177.0   3rd Qu.: 95.30
                                       Max.   :208.3   Max.   :176.40
     sysBP            diasBP             hr           O2_hours_day
 Min.   : -1.0   Min.   : -1.00   Min.   : -1.00   Min.   : 0.0000
 1st Qu.:118.0   1st Qu.: 70.00   1st Qu.: 65.00   1st Qu.: 0.0000
 Median :128.0   Median : 77.00   Median : 73.00   Median : 0.0000
 Mean   :128.7   Mean   : 76.81   Mean   : 74.15   Mean   : 0.9468
 3rd Qu.:140.0   3rd Qu.: 84.00   3rd Qu.: 82.00   3rd Qu.: 0.0000
 Max.   :211.0   Max.   :120.00   Max.   :131.00   Max.   :24.0000
      bmi            asthma          hay_fever      bronchitis_attack
 Min.   :12.67   Length:5747        Min.   :0.0000   Length:5747
 1st Qu.:24.68   Class :character   1st Qu.:0.0000   Class :character
 Median :28.20   Mode  :character   Median :0.0000   Mode  :character
 Mean   :29.08                      Mean   :0.5243
 3rd Qu.:32.42                      3rd Qu.:1.0000
 Max.   :64.10                      Max.   :3.0000
  pneumonia          chronic_bronchitis  emphysema            copd
 Length:5747        Length:5747        Length:5747        Length:5747
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


 sleep_apnea          SmokStartAge   CigPerDaySmokAvg Duration_Smoking
 Length:5747        Min.   :-1.00   Min.   :-1.00     Min.   :-1.00
 Class :character   1st Qu.:14.00   1st Qu.:20.00     1st Qu.:29.80
 Mode  :character   Median :16.00   Median :20.00     Median :36.00
                    Mean   :16.73   Mean   :23.76     Mean   :35.02
                    3rd Qu.:18.50   3rd Qu.:30.00     3rd Qu.:42.00
                    Max.   :50.00   Max.   :99.00     Max.   :67.00
 smoking_status     total_lung_capacity pct_emphysema
 Length:5747        Min.   :-1.000      Min.   :-1.0000
 Class :character   1st Qu.: 3.722      1st Qu.: 0.1644
```

Multiple Imputation:

As the size of the dataset and the missing values are large in number , it is best practice to do multiple imputation of data for all NA's in the COPD dataset.

Missing values imputation

```
[ ] copd_imputed_data <- mice(copd, method = 'pmm', m = 5)
    completed_data <- complete(copd_imputed_data, 1)
```

```
 iter imp variable
  1   1  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  1   2  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  1   3  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  1   4  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  1   5  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  2   1  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  2   2  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  2   3  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  2   4  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  2   5  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  3   1  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  3   2  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  3   3  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  3   4  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  3   5  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  4   1  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  4   2  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  4   3  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  4   4  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  4   5  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  5   1  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  5   2  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  5   3  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  5   4  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
  5   5  sysBP  diasBP  hr  SmokStartAge  CigPerDaySmokAvg  Duration_Smoking  total_lung_capacity  pct_emphysema  functional_residual_capacity  pct_gastrapping  insp_meanatt  exp_meanatt  FEV1_FVC_ratio  FEV1  FVC
Warning message:
"Number of logged events: 12"
```

```
# to narrow down more
options(repr.plot.width=20, repr.plot.height=10)

# removing non-numeric and ID/date columns
copd2 <- na.omit(dplyr::select(copd, -sid, -visit_year, -visit_date))

# Check for multicollinearity
cor <- cor(select_if(copd2, is.numeric))
high_cor <- which(abs(cor) > 0.70, arr.ind = TRUE)

# Now create your predictors and response for the regression subset selection
response <- copd2$pct_emphysema
predictors <- model.matrix(pct_emphysema ~ ., data = copd2)[,-1]

# Run best subsets selection
best_subsets <- regsubsets(predictors, y = response, nbest = 1, nvmax = NULL, really.big = TRUE)

# Plot the results
plot(best_subsets, scale = "bic")
```

```
[ ] # serial correlation test
    durbinWatsonTest(fit_final_clean)
```

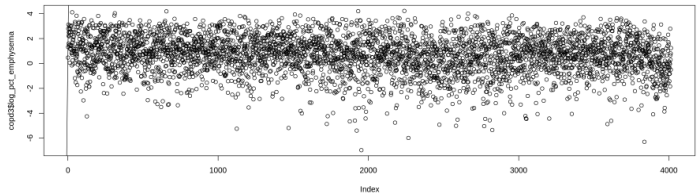
```
 lag Autocorrelation D-W Statistic p-value
   1       0.03599563      1.927705   0.016
 Alternative hypothesis: rho != 0
```

```
# serial correlation test
durbinWatsonTest(fit_interaction1)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.03554443      1.928598   0.012
 Alternative hypothesis: rho != 0
```

```
[ ]
```

```
[ ] plot(predict(fit_interaction1), copd3$log_pct_emphysema)
    abline(0, 1)
```



Explanation of the Variables:

Variable	Description
sid	Anonymized patient ID
visit_year	Year of visit
visit_date	Data of visit
visit_age	Age at time of visit
gender	Gender (Male, Female)
race	Race (White, Black or African American)
height_cm	Height in centimeters
weight_kg	Weight in kilograms
sysBP	Systolic blood pressure
diasBP	Diastolic blood pressure
hr	Heart rate
O2_hours_day	On typical 24-hour day, how many hours supplemental 02 used [hours]
bmi	Body mass index
asthma	Have you ever had asthma? (Yes, No, unknown)
hay_fever	Have you ever had hay fever? (Yes, No, unknown)
bronchitis_attack	Have you ever had a bronchitis attack? (Yes, No, unknown)
pneumonia	Have you ever had pneumonia? (Yes, No, unknown)
chronic_bronchitis	Have you ever had chronic bronchitis? (Yes, No, unknown)
emphysema	Have you ever had emphysema diagnosed by a clinician? (Yes, No, unknown)
copd	Have you ever had copd diagnosed by a clinician? (Yes, No, unknown)
sleep_apnea	Have you ever had sleep apnea? (Yes, No, unknown)
SmokStartAge	How old were you when you first started cigarette smoking? (years old)
CigPerDaySmokAvg	Average for entire time, how many cigarettes smoked per day (cigarettes/day)
Duration_Smoking	Duration of smoking, years
smoking_status	Never-smoked, Former smoker, Current smoker
total_lung_capacity	Volume of air in lungs at full inspiration (full breath-hold) [Liters]
pct_emphysema	Percentage of emphysema (damaged lung areas) [%]
functional_residual_capacity	Volume of air in lungs at expiration (exhale) [Liters]
pct_gastrapping	Percentage of air trapping in lungs after exhaling [%]
insp_meanatt	Average lung density at full inspiration (full breath-hold) [Hounsfield units]
exp_meanatt	Average lung density at expiration (exhale) [Hounsfield units]
FEV1_FVC_ratio	Ratio between FEV1 and FVC
FEV1	Forced expiratory volume in 1 second - volume of air forcefully exhaled in 1 second
FVC	Forced vital capacity - volume of air exhaled after full breath
FEV1_phase2	FEV1 five years later - volume of air forcefully exhaled in 1 second

Statistical Inferences Plots:

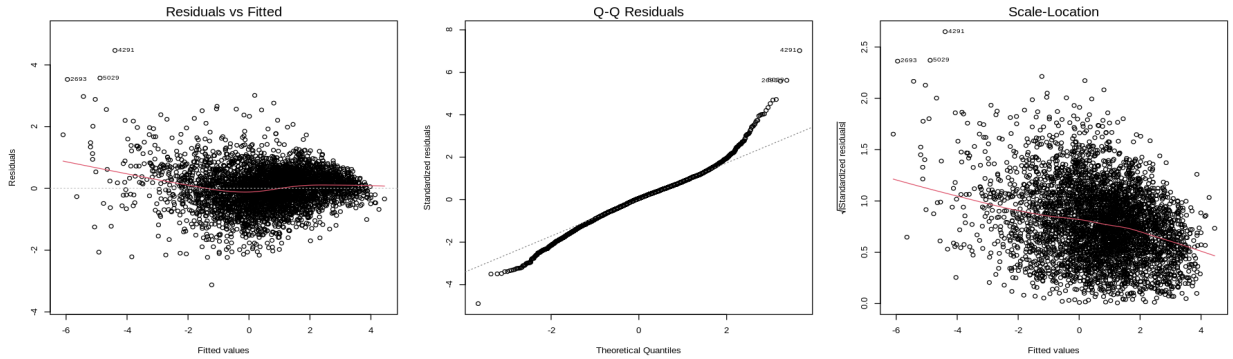


Fig 1b: Residual Diagnostics for COPD Dataset Linear Regression Model without transformation

The scatter plot shows the relationship between actual and predicted percentages of emphysema, with the predicted values derived from a ridge regression model; the plot highlights the model's predictive performance across the range of observed data.

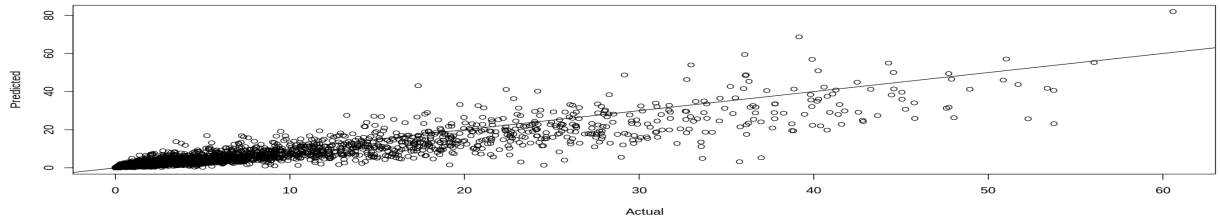


Fig 2b: Actual vs Predicted Emphysema Percentage from Ridge Regression Model

The output displays a summary of the regression coefficients from a statistical model, detailing the estimates, standard errors, t-values, and p-values, indicating the significance of predictors such as oxygen hours per day, smoking status, and various lung function measures on the log-transformed percentage of emphysema.

```
t test of coefficients:

                                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                         -1.6118e+01  6.9911e-01 -23.0547 < 2.2e-16
O2_hours_day                         1.0995e-02  2.5458e-03   4.3189 1.606e-05
smoking_statusFormer smoker          6.9464e-02  2.3110e-02   3.0058 0.0026649
log(total_lung_capacity)            -1.4837e+00  2.1764e-01  -6.8173 1.067e-11
log(functional_residual_capacity)    1.3831e+00  1.9171e-01   7.2146 6.441e-13
log(pct_gastrapping)                 1.2404e+00  2.9496e-02  42.0515 < 2.2e-16
insp_meanatt                        -3.9254e-02  1.3778e-03 -28.4904 < 2.2e-16
exp_meanatt                          2.1823e-02  8.7824e-04  24.8487 < 2.2e-16
FEV1_FVC_ratio                      -3.8139e+00  2.8192e-01 -13.5284 < 2.2e-16
FEV1                                 3.6935e-01  8.2606e-02   4.4712 7.992e-06
FVC                                 -1.8475e-01  4.6660e-02  -3.9596 7.639e-05
FEV1_phase2                         -1.4150e-01  3.7224e-02  -3.8012 0.0001462

(Intercept)                         ***
O2_hours_day                        ***
smoking_statusFormer smoker         **
log(total_lung_capacity)            ***
log(functional_residual_capacity)   ***
log(pct_gastrapping)                ***
insp_meanatt                        ***
exp_meanatt                         ***
FEV1_FVC_ratio                      ***
FEV1                                ***
FVC                                 ***
FEV1_phase2                         ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 3b: Regression Coefficients for Predicting Log-Transformed Percentage of Emphysema

+----------+-----+----------+---------+-------------+------------------+

Factor	Df	Sum Sq	Mean Sq	F value	P-Value
visit_age	1	1232.46	1232.46	2551.15	<0.0001
smoking_status	1	432.77	432.77	895.82	<0.0001
lung_function_index	1	2727.96	2727.96	5646.80	<0.0001
log(pct_gastrapping)	1	3122.76	3122.76	6464.03	<0.0001
meanatt_diff	1	1049.58	1049.58	2172.60	<0.0001
bmi	1	85.31	85.31	176.59	<0.0001
FEV1_FVC_ratio	1	618.79	618.79	1280.88	<0.0001
FEV1_phase2	1	40.58	40.58	84.00	<0.0001
Residuals	4004	1934.33	0.48	NA	NA

Table 1b: ANOVA Results for Regression Model Predicting COPD Progression

The data frame outlines three influential observations from a regression model, detailing their standardized residuals (StudRes), leverage (Hat), and Cook's Distance (CookD), with observation 2693 showing particularly high leverage and influence on the model fit.

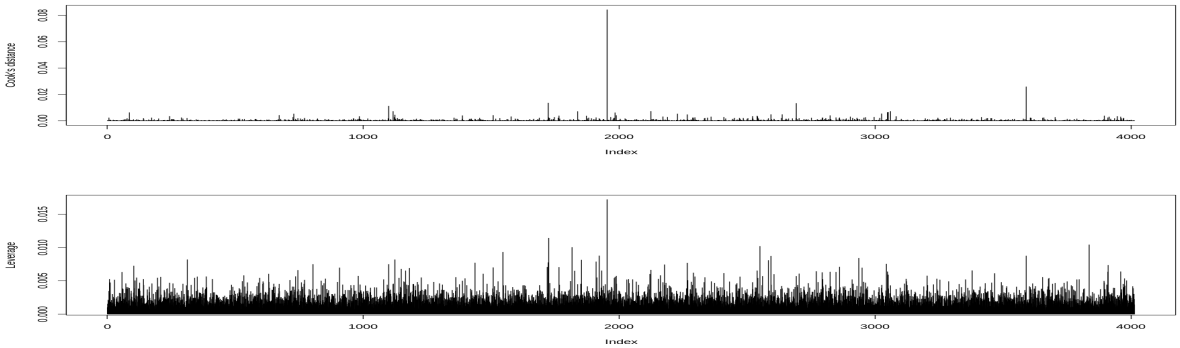


Fig 4b: Analysis identifies observations 2343, 2693, and 5029 as influential, with 2693 exerting the most substantial leverage and impact according to Cook's Distance, potentially warranting further investigation

The triptych of plots displays Cook's distance, residuals vs. leverage, and Cook's distance vs leverage plot influence measures for a regression model, highlighting points 2343, 2693, and 5029 as influential cases that may disproportionately affect the model's predictions.

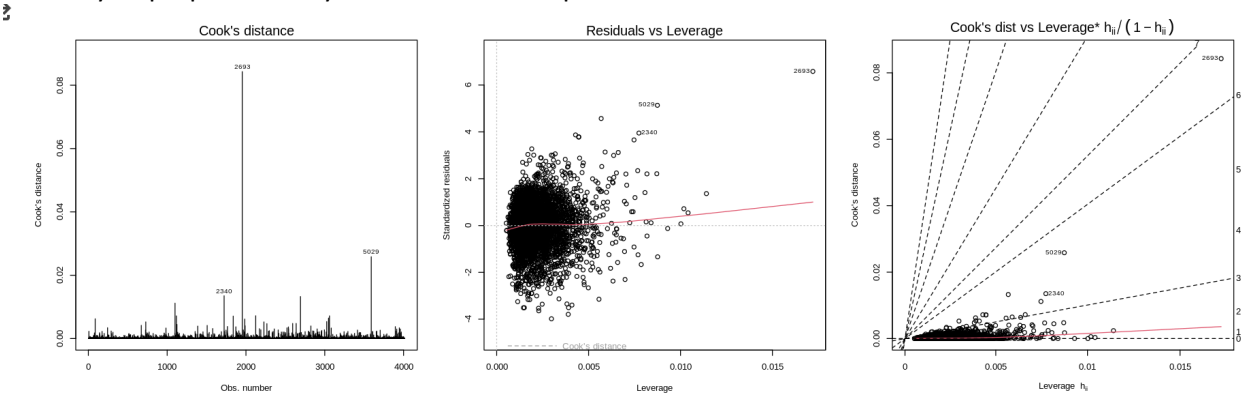


Fig 5b: Influence Diagnostics for Regression Model: Cook's Distance and Leverage Analysis

The data represent variance inflation factor (VIF) values for each predictor in a regression model, indicating the degree of multicollinearity, with variables like log(pct_gastrapping) and FEV1_FVC_ratio showing high VIFs, suggesting strong multicollinearity.

visit_age: 1.53234587287561 smoking_status: 1.36888719614381 lung_function_index: 2.58880201304239 log(pct_gastrapping): 3.7844578895866 meanatt_diff: 3.69756548595131 bmi: 1.12830000872957 FEV1_FVC_ratio: 4.14804611535346 FEV1_phase2: 3.46110389487486

Fig 6b: Variance Inflation Factor (VIF) for Assessing Multicollinearity in COPD Regression Model

This table presents the estimated coefficients from a regression analysis, showcasing the relationship between various predictors and the outcome variable. Significance levels are indicated, with all variables showing strong statistical significance ($p < 0.001$).

Variable	Estimate	SE	t value	p-value
(Intercept)	-2.841	0.168	-16.933	<0.001
visit_age	-0.014	0.002	-8.963	<0.001
smoking_statusFormer smoker	0.301	0.025	11.653	<0.001
lung_function_index	0.179	0.015	11.494	<0.001
log(pct_gastrapping)	1.567	0.019	82.750	<0.001
meantt_diff	-0.562	0.004	56.746	<0.001
bmi	-0.028	0.002	-14.482	<0.001
FEV1_FVC_ratio	-0.310	0.153	-20.305	<0.001
FEV1_phase2	-0.219	0.024	-9.086	<0.001

Table 1b. Statistical Output

These plots collectively indicate that the assumptions of linear regression (linearity, normality of errors, and homoscedasticity) may not be fully satisfied, potentially affecting the validity of model inferences. Consider investigating further with transformations, robust methods, or alternative models.

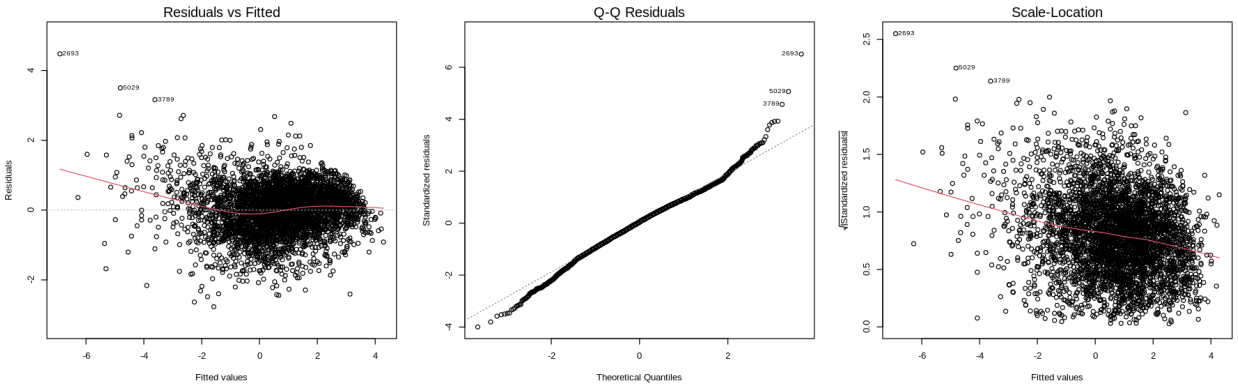


Fig 7b: Analysis of Fina Model Diagnostics plots

Since most of the data points are below the threshold, the overall influence on the regression model is likely minimal for the majority of the dataset.

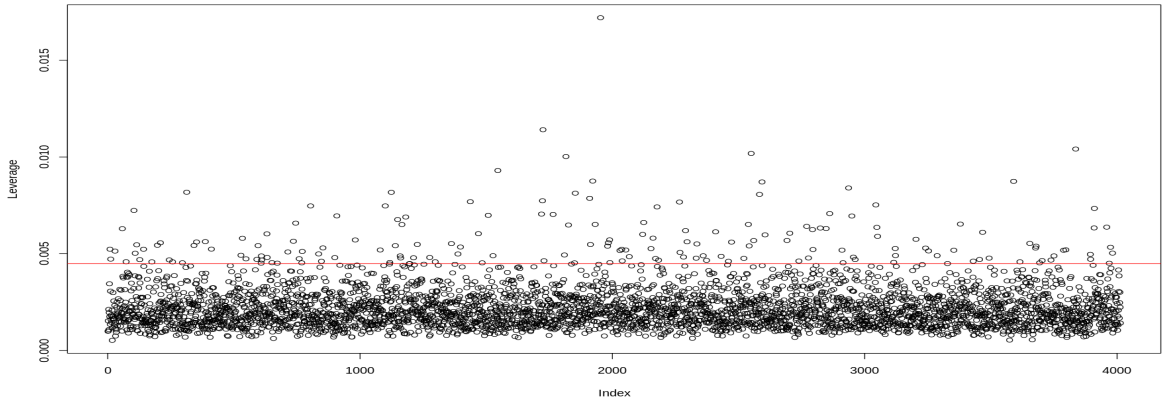


Fig 8b: Leverage Plot for Regression Model Diagnostics

These plots are generally used to identify the need for potential model refinements, such as the inclusion of interaction terms, or to validate the inclusion of each predictor in the model. The plots suggest that while the model may capture the linear relationships well, further investigation into outliers and influential points is warranted.

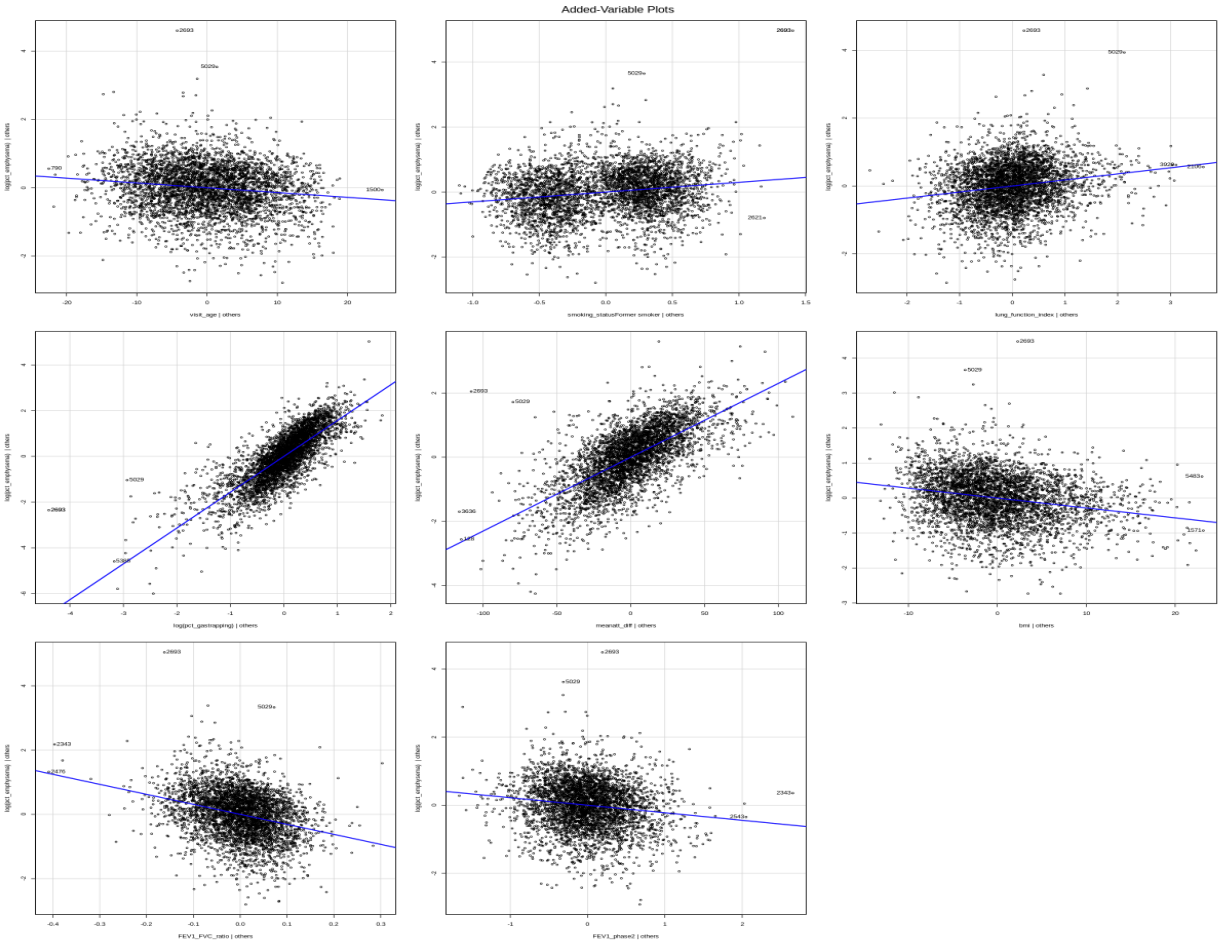


Fig9b. Partial Regression Diagnostic Plots for Multiple Regression Analysis