**"Investigating Emphysema Severity in COPD: A Regression Approach"**

**Clinical analysis of emphysema**

SAI TEJASRI YERRAMSETTI

**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.

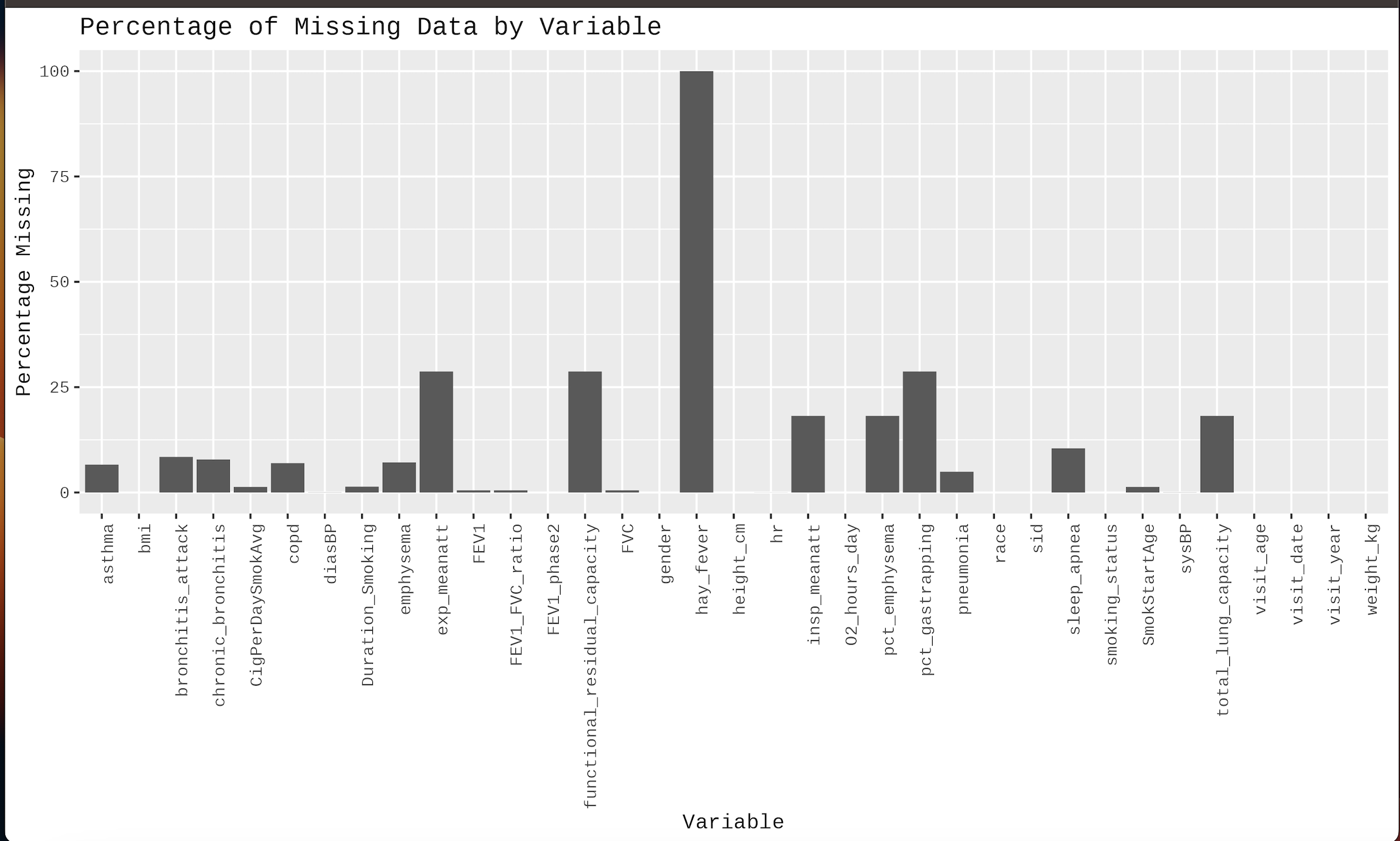
**Exploratory Data Analysis**

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**

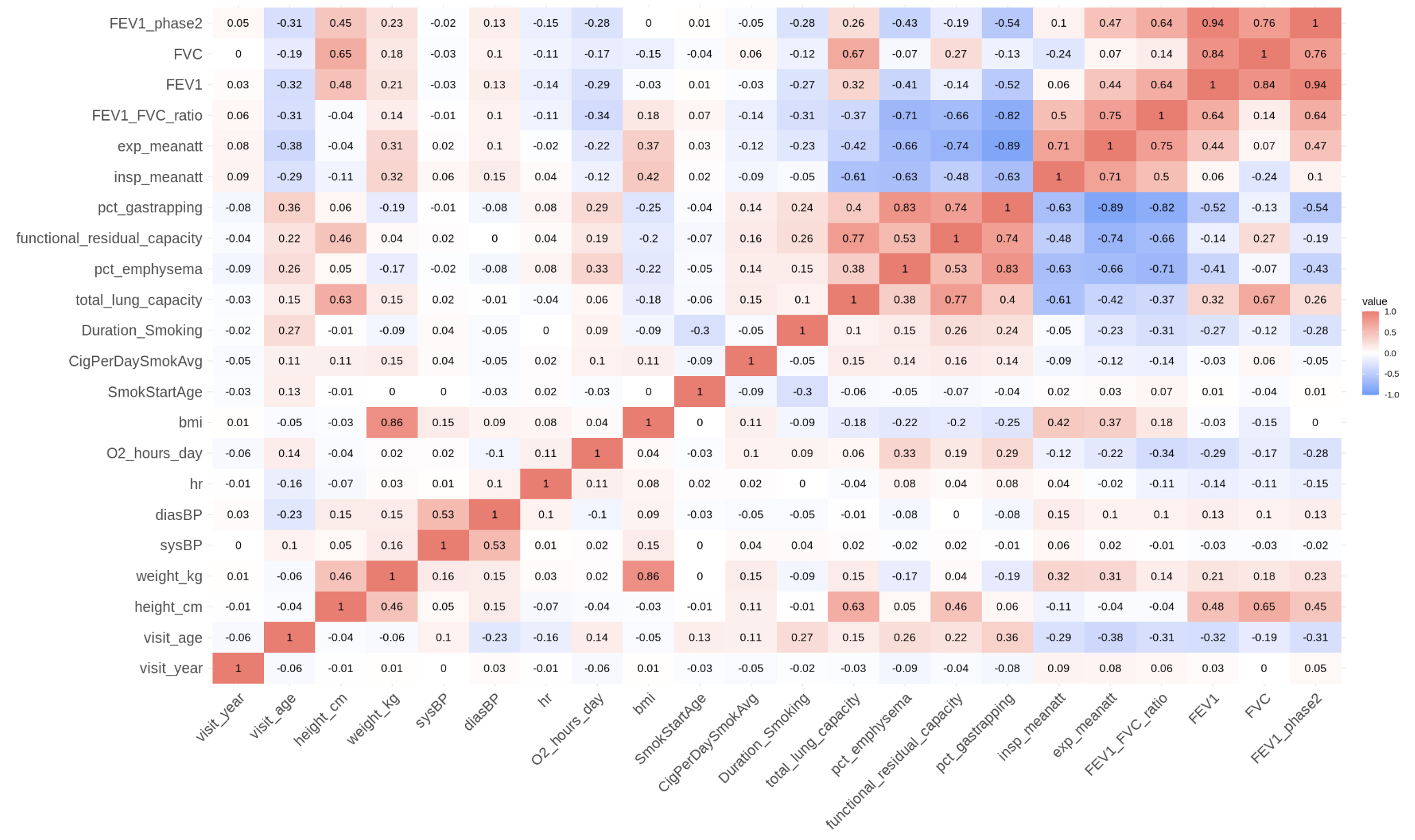
The summary statistics provided indicate a dataset with 5747 observations, encompassing variables related to demographic, clinical, and lifestyle factors of COPD patients from the COPDGene study. Key variables show a wide range of values, such as age (39 to 85 years), BMI (12.67 to 64.10), and total 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.

****

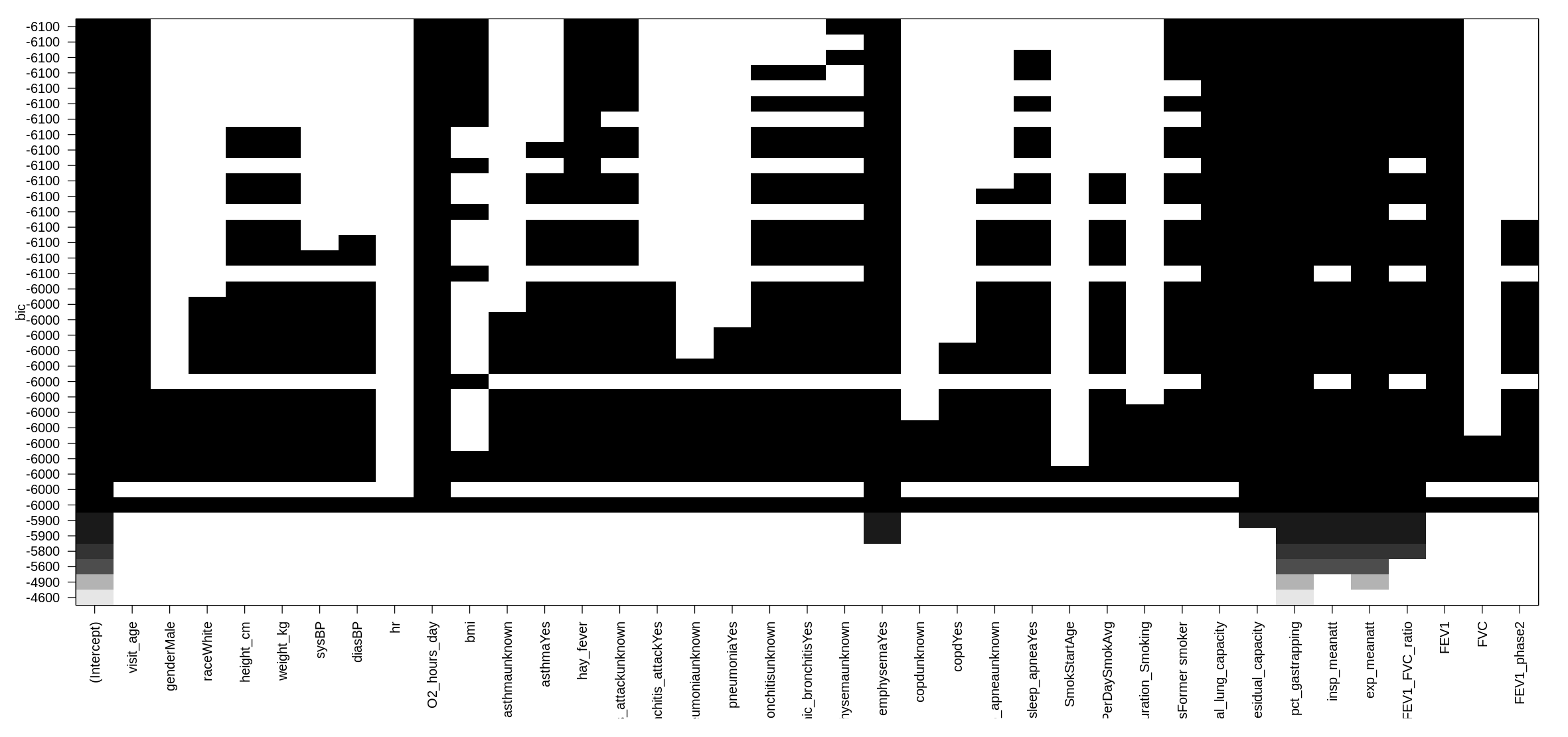
**Fig 1a: Bar Chart for Percentage of Missing Data by Variable**

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 (H₀):

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 (H₁):

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(pct\_emphysema)=β0+β1(visit\_age)+β2(smoking\_status)+β3(total\_lung\_capacity)+

β4(log(pct\_gastrapping))+β5(meanatt\_diff)+β6(bmi)+β 7(FEV1\_FVC\_ratio)+β 8(FEV1\_phase2)+ϵ

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

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

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

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

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

# 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.

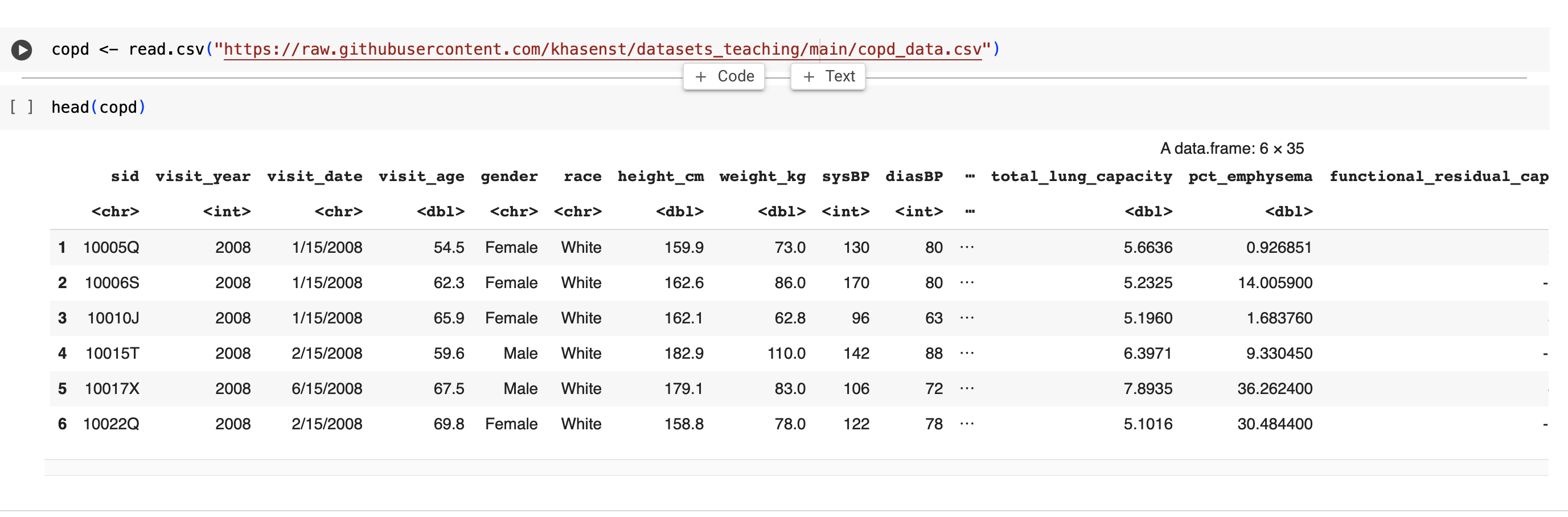
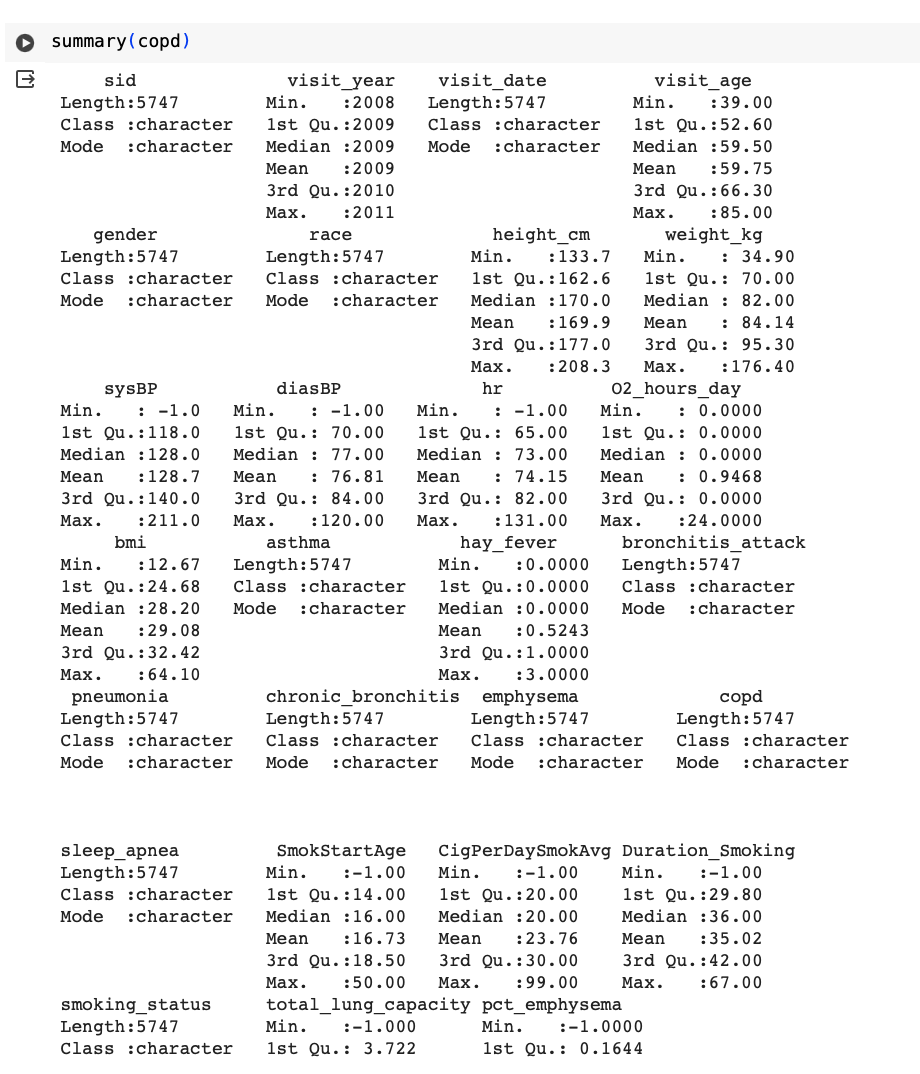
# Cook, D. R. "Detection of influential observations in linear regression." Technometrics, vol. 19, pp. 15-18, 1977.

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

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

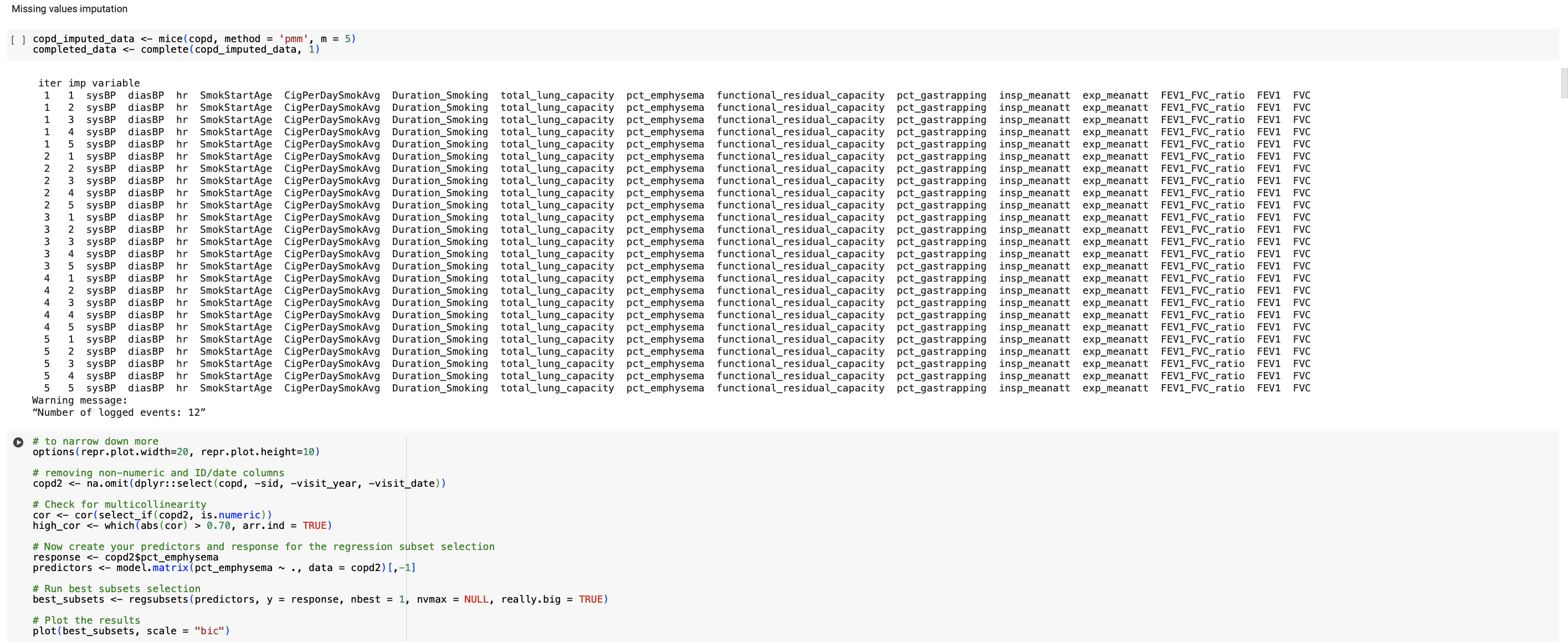
**Appendix**

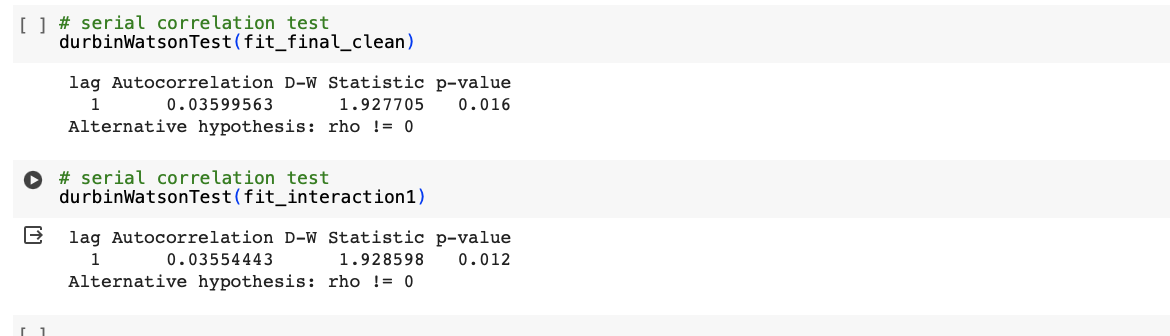
We’ll begin the appendix with an annotated version of the code used to produce the outputs in this report

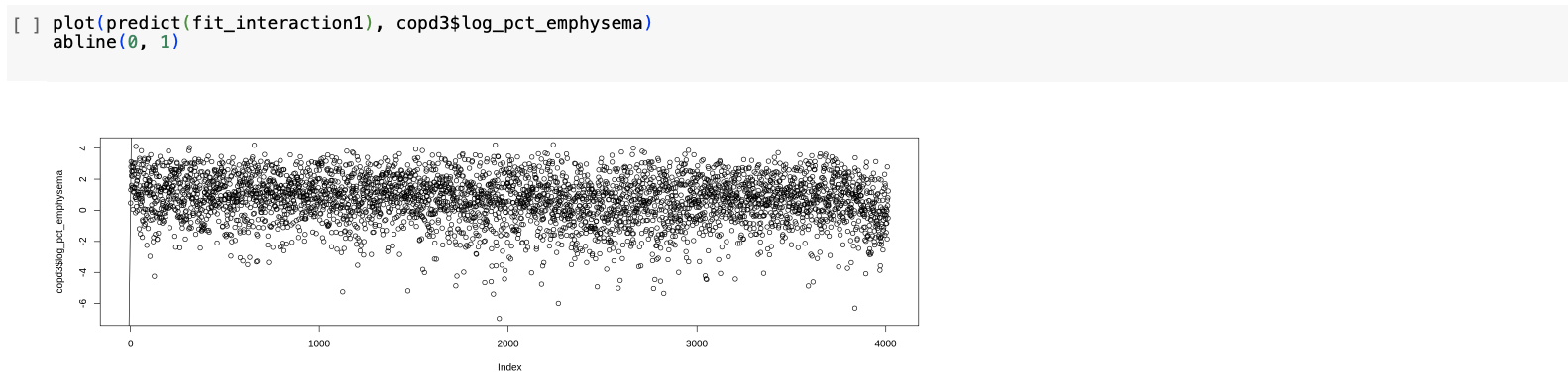
 

Multiple Imputation:

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.



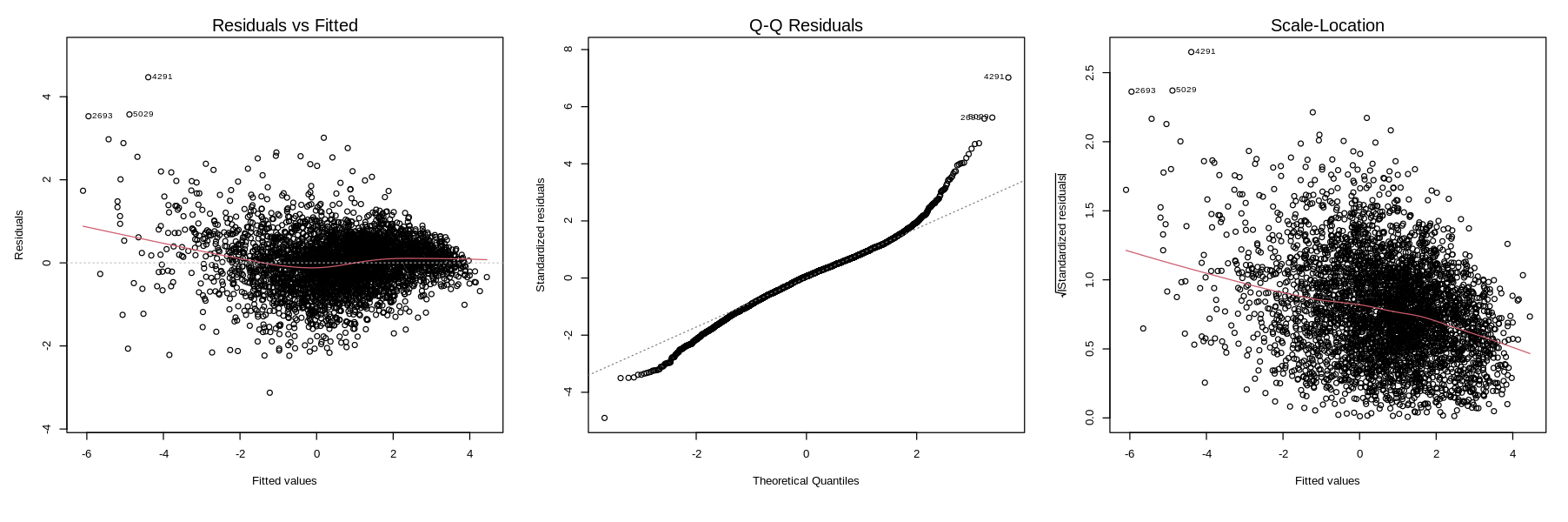




**Explanation of the Variables:**

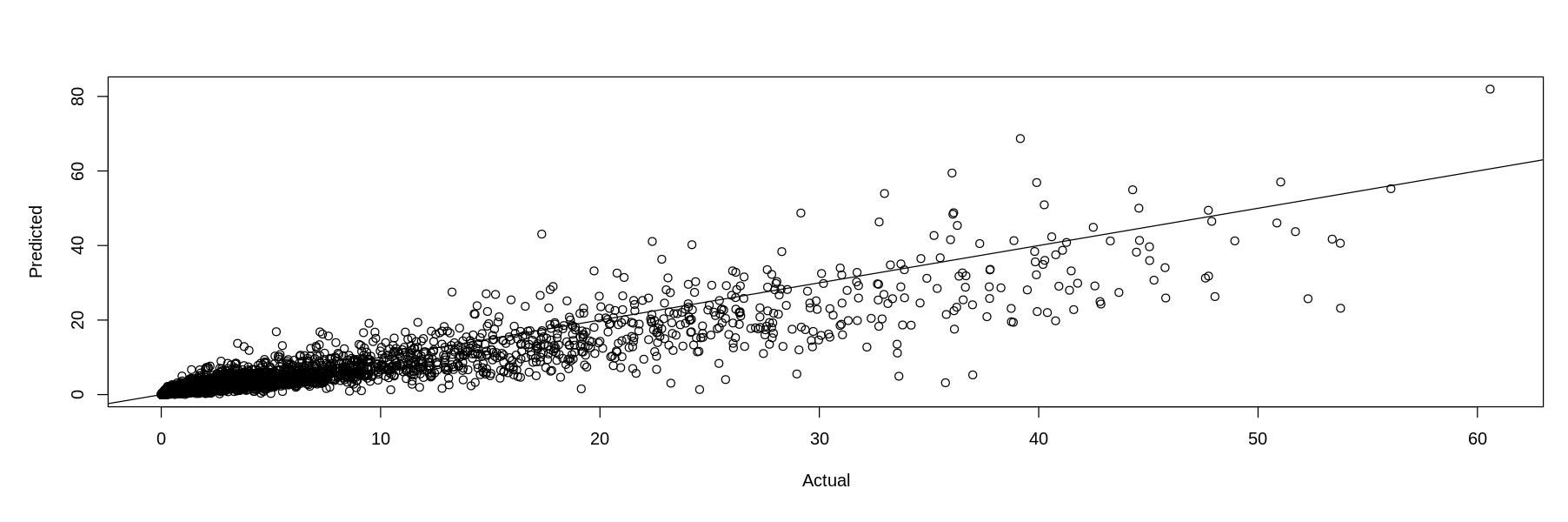
****

**Statistical Inferences Plots:**

****

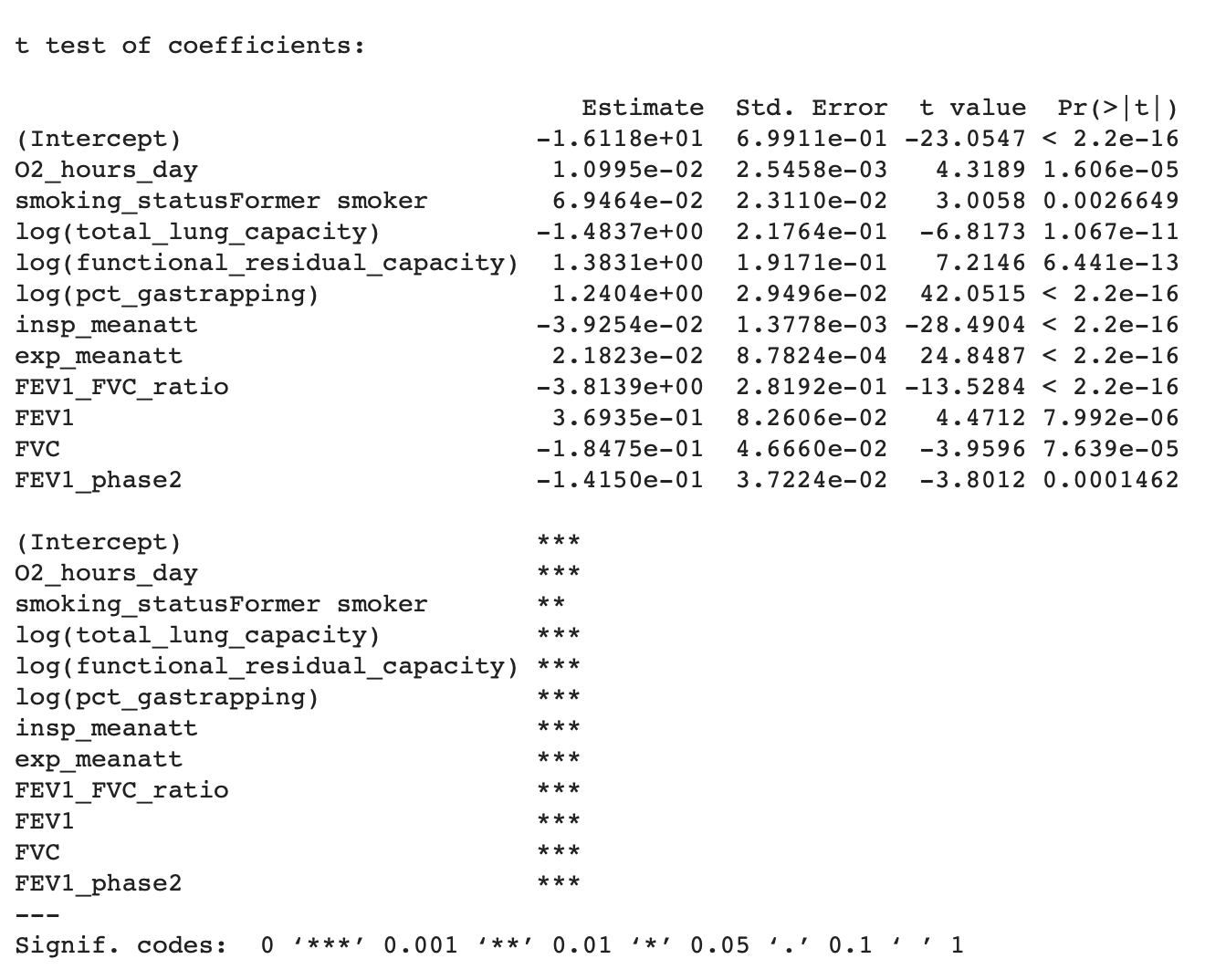
**Fig 1b: Residual Diagnostics for COPD Dataset Linear Regression Model without transformation**

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.

****

**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 |**

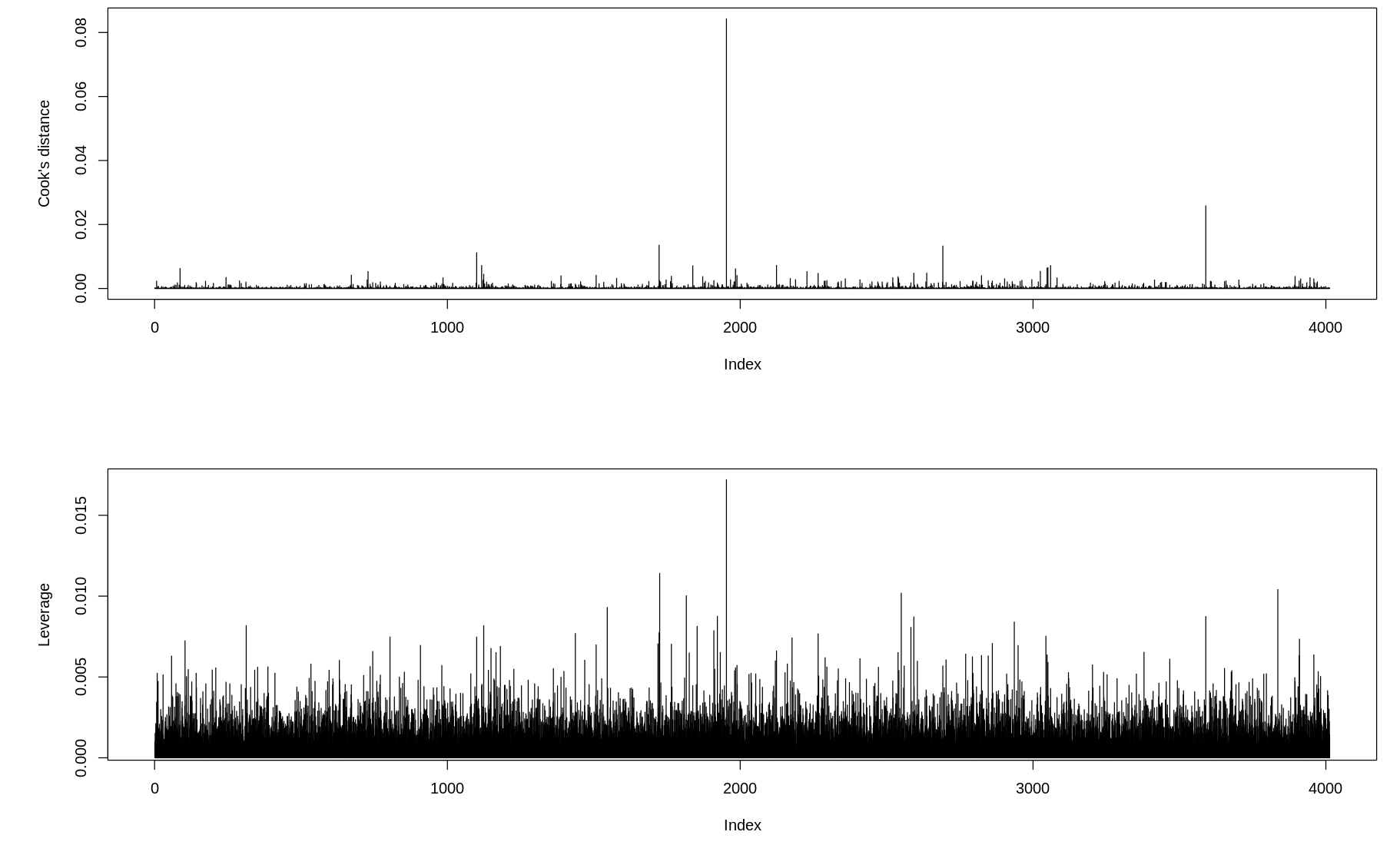
**| FEV1\_phase2 | 1 | 40.58 | 40.58 | 84.00 | <0.0001 |**

**| Residuals | 4004| 1934.33 | 0.48 | NA | NA |**

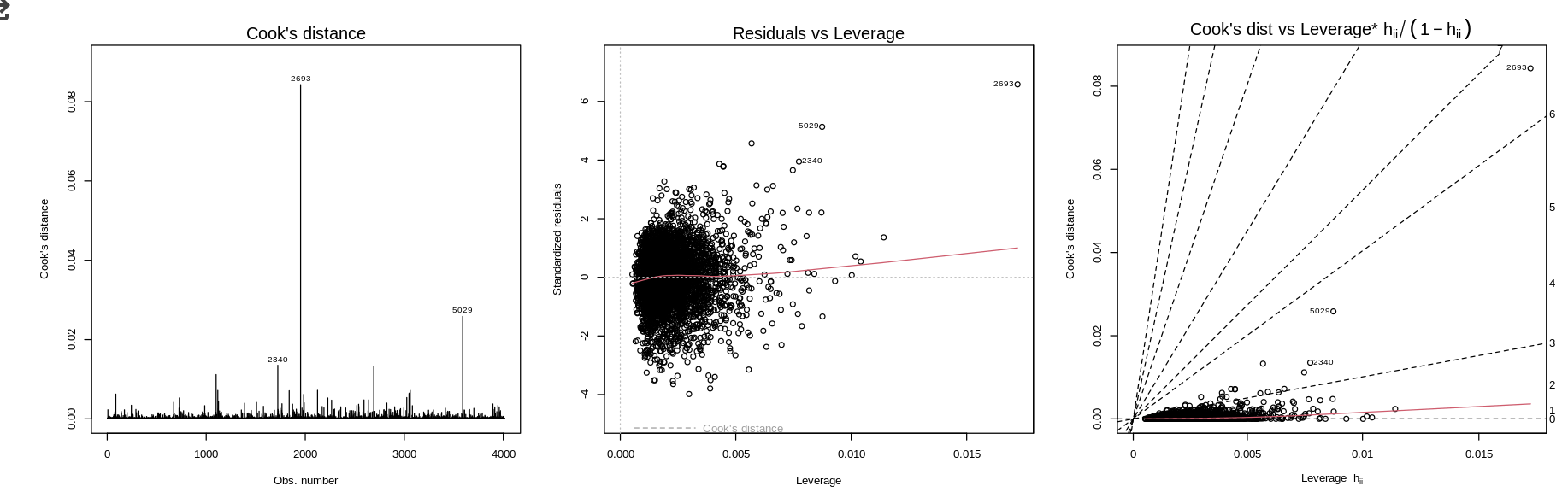
**+------------------------+-----+-----------+---------+-------------+---------------------+-**

**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.



**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**

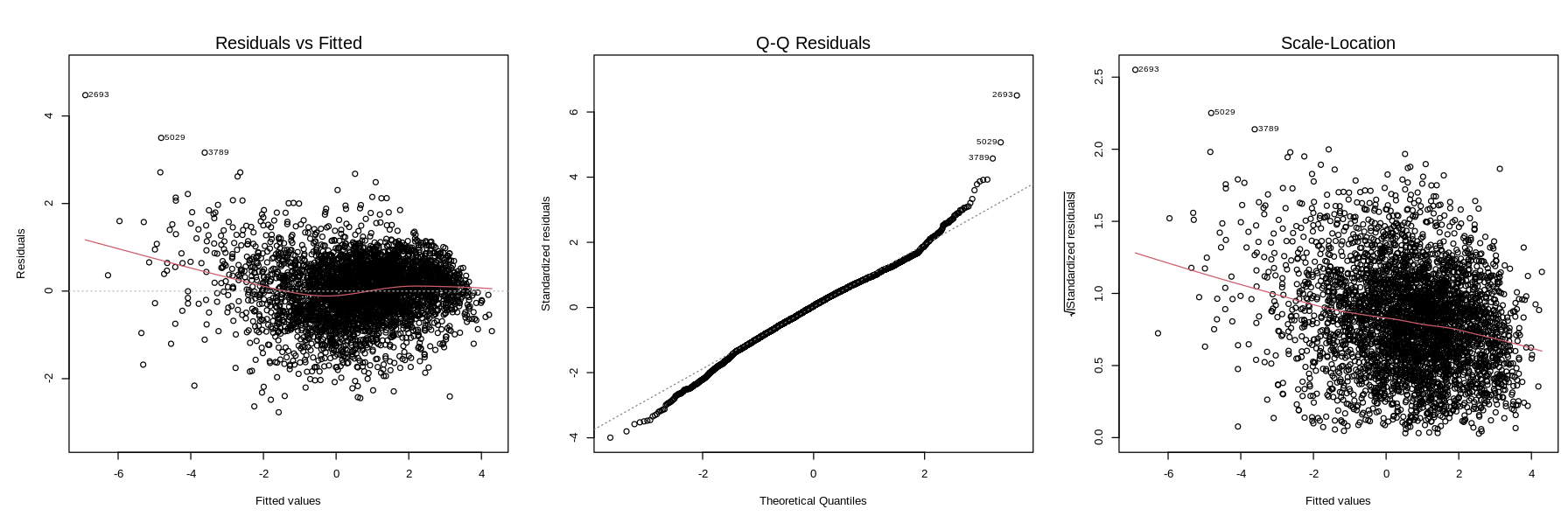
**bmi -0.028 0.002 -14.482 <0.001**

**FEV1\_FVC\_ratio -0.310 0.153 -20.305 <0.001**

**FEV1\_phase2 -0.219 0.024 -9.086 <0.001**

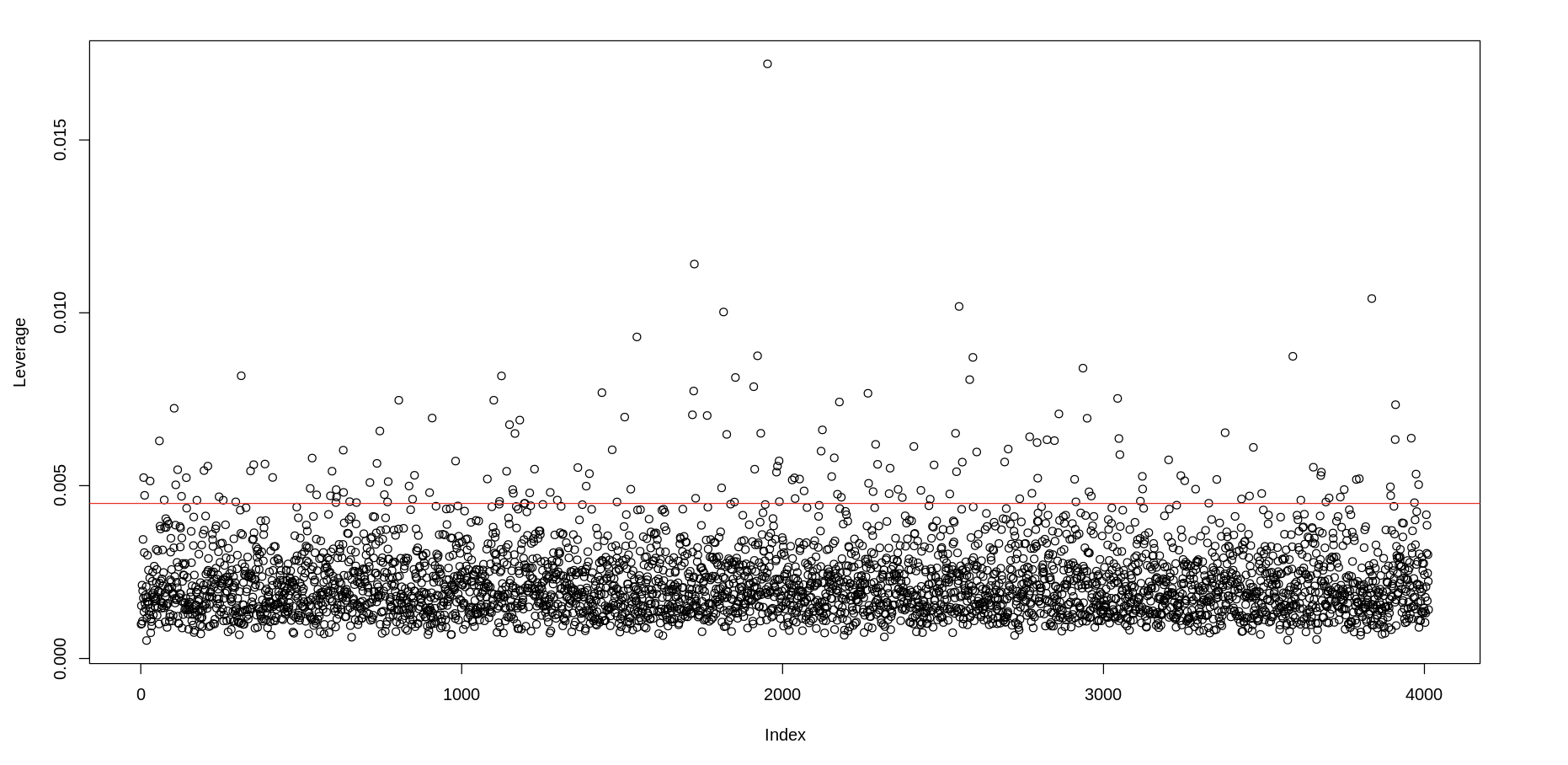
**Table 1b. Statistical Output**

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.

****

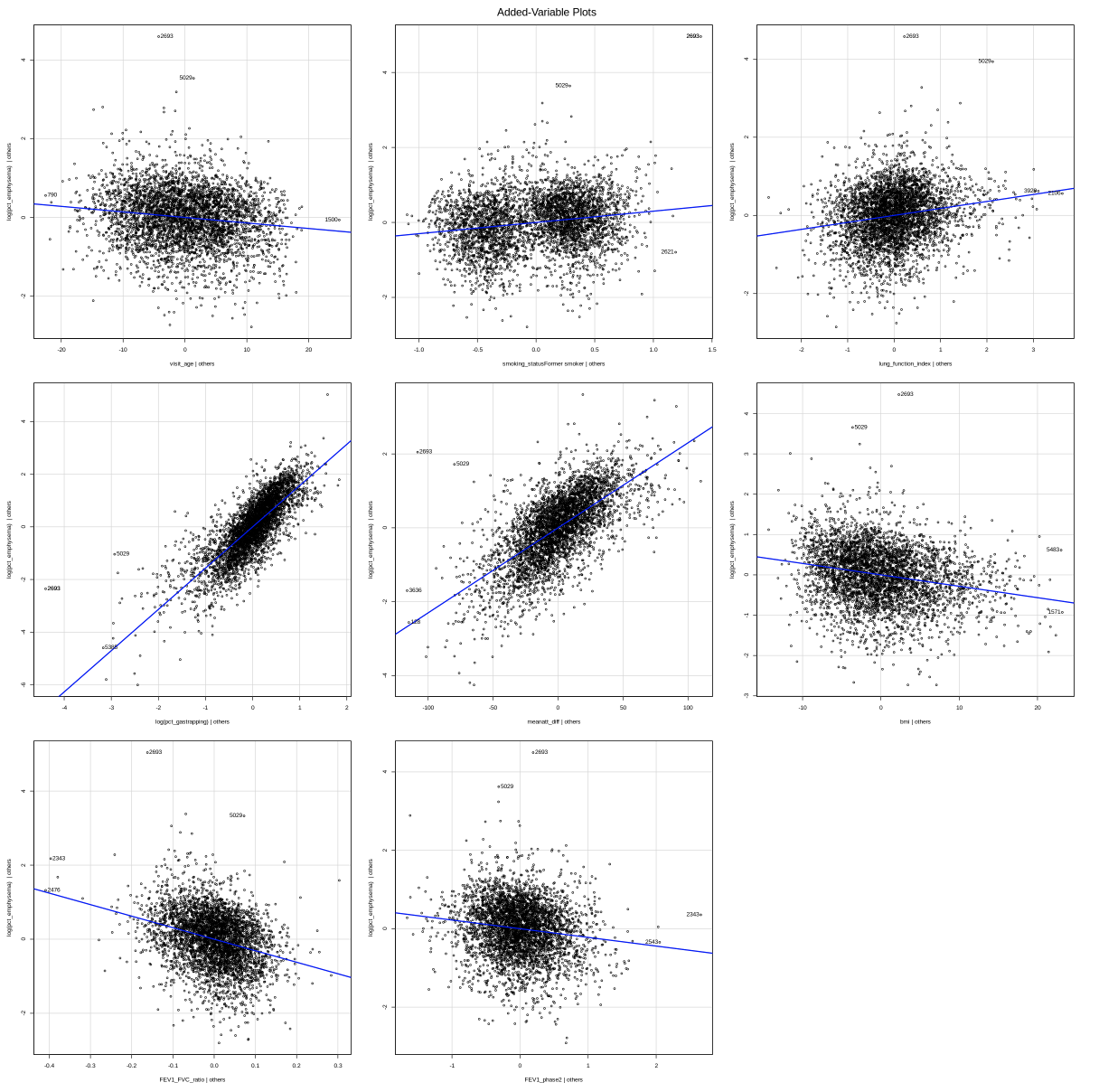
**Fig 7b: Analysis of Fina Model Diagnostics plots**

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**